MedExpDial: Machine-to-Machine Generation of Explanatory Dialogues for Medical QA

Andrea Zaninello
Fondazione Bruno Kessler
Free University of Bolzano (Italy)
azaninello@fbk.eu

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

1 Motivations and Background

We describe a pilot study on generating synthetic explanatory dialogues for the medical domain, based on a pre-existing medical dataset of multiple-choice questions with human-written explanations. We use an instruction-tuned large language model (LLM) to generate dialogues between a medical student and a teacher/doctor helping answer questions about clinical cases. We inject varying degrees of background knowledge into the teacher prompt and analyze the effectiveness of these dialogues in terms of whether the student is able to get to the correct answer and in how many turns. This method has potential applications in developing and evaluating argument-based explanation models for medical question answering (QA).

Currently, medical QA systems and health-related AI systems are increasingly being used to provide patients with access to reliable information, support healthcare professionals in their decision-making processes, or for educational purposes (Kell et al., 2024; Alonso et al., 2024; Yagnik et al., 2024; García-Ferrero et al., 2024). A key challenge in this field is providing explanations that are both accurate and understandable to the user (Li'evin et al., 2022), as they play a crucial role in building trust and transparency in AI systems, particularly in critical domains like healthcare (Hossain et al., 2023).

On the one hand, traditional approaches to explanation generation in medical QA often involve providing static summaries, rule-based or template-based explanations (Budler et al., 2023). However, these approaches are only partially able to capture the reasoning involved in medical diagnosis and treatment (Li'evin et al., 2022; Molinet et al., 2024). On the other hand, by engaging the users in a conversation, dialogue systems can provide more interactive explanations, adapting to the user's specific needs and understanding, which can be dynamically tailored through interactions and feedback in a dialogue flow (Wachsmuth and Alshomary, 2022). However, because of the highly sensitive nature of medical records, ecological data are extremely difficult to collect in this domain.

```xml
&lt;?xml version="1.0" ?&gt;
&lt;full_question id="23_113" type="INFECTIOUS"&gt;
  A 71-year-old woman with a history of rheumatoid arthritis on sulfasalazine,
prednisone and etanercept. She goes to the emergency room for 72 hours of clinical
manifestations compatible with facial herpes zoster affecting the right hemiface,
auricular pavilion, respecting the forehead and conjunctival chemosis. What would be
the appropriate treatment?
  &lt;full_answer&gt;I think this question is not clearly in the Infectious Diseases
syllabus, and may overlap with OFT and DERMA, but as I understand it, in an
immunocompromised patient and also with data of ocular involvement, admission for
intravenous treatment would be indicated due to the high risk of possible
complications.&lt;/full_answer&gt;
  &lt;option num="1"&gt;Symptomatic treatment of pain only.&lt;/option&gt;
  &lt;option num="2"&gt;Topical treatment with acyclovir.&lt;/option&gt;
  &lt;option num="3"&gt;Outpatient treatment with acyclovir, valacyclovir or oral
famciclovir.&lt;/option&gt;
  &lt;option num="4" correct="YES"&gt;
    Hospital admission and treatment with acyclovir or famciclovir iv.
    &lt;explanation char_ranges="[[135, 310]]" word_ranges="[[24, 50]]"&gt;in an
immunocompromised patient and also with data of ocular involvement, admission for
intravenous treatment would be indicated due to the high risk of possible
complications.&lt;/explanation&gt;
  &lt;/option&gt;
  &lt;option num="5"&gt;Parenteral Ig and vaccination.&lt;/option&gt;
&lt;/full_question&gt;
```

Figure 1: An xml-coded question, answers and explanations from the CasiMedicos dataset.

To fill this gap, we explore the generation of dialogue-based medical explanations in an educational setting (Anonymous, 2024), as a way to enhance the explainability of medical QA systems, contributing to developing effective medical dialogue models.

2 Explanatory Dialogue Generation

Our explanatory dialogues are based on *CasiMedicos*, a pre-existing dataset of medical questions and answers with human-written explanations (Agerri et al., 2023), which contains questions in Spanish, English, French, Basque, and Italian, covering various medical specialties. Every language corresponds to a train, test, dev splits of 434, 125, and 63 questions each. Each question consists of a clinical case followed by a question on the case, 5 multiple-choice options of which one is the correct answer, and a human-written explanation for the correct answer and/or for the reason why the other

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

options are not correct. An example question from *CasiMedicos* is provided in Figure 1.

The first step is to identify the questions in CasiMedicos that a state-of-art LLM is *unable* to correctly answer, under the assumption that its internal knowledge alone is not sufficient to answer them. To do this, we prompt an instance of GPT-4 (OpenAI, 2023) to answer the 125 questions of the English split of the CasiMedicos test set, without any help (0-shot). We parse the model's answers with regular expressions and compare them with the CasiMedicos correct answers. GPT-4 was able to answer 105 over 125 questions correctly, yelding an initial accuracy of 84%.

Then, we use the 20 answers that the model was unable to answer correctly and two independent instances of GPT-4, a medical *Teacher* and a medical *Student*, to generate dialogues. The Teacher is prompted to help a student prepare for the USMLE exam, and incrementally provided with more information from the knowledge base, while the Student is only prompted to play the role of the student with no additional information[1].

We experiment with four different modes of dialogue generation corresponding to the information provided to the Teacher instance. Specifically, the Teacher is only provided with the clinical case without the correct answer (More 0), or incrementally with the correct answer (Mode 1), the alternative options (Mode 2), and the human-written explanation (Mode 3).

The Teacher is allowed to use any of the provided information as she wishes to guide the conversation and help the Student reach the correct answer. The Teacher is also prompted to end the conversation when the final answer is reached, outputting an <END> tag once the Student identifies the correct answer. For each question, 2 different dialogues are generated for each mode, ranging from a minimum of 6 turns to a maximum of 10 turns, for a total of 160 dialogues. We split the generated dialogues into an 80-dialogue test and dev sets.

Finally, students from the University of Bologna manually annotated each dialogue of the test set for the following elements: 1. *Answer Detection*, i.e., the text fragment within the dialogue where the Student provides her final answer; 2. *Option Mapping*: a mapping between the Student's final answer and the original question's option[2]; 3. *Answer Correctness*: whether the Student's answer is correct based on the knowledge base. We manually and semi-automatically revise the annotation and evaluate the effectiveness of the dialogues in the different modes by measuring the accuracy of each dialogue mode as well as the number of turns it takes for the Student to get to the correct answer. A lower number of turns should in fact indicate a more effective dialogue.

Mode	Correct	Accuracy	Mean Turns
Mode 0	9	0.45	4.5
Mode 1	13	0.65	5.1
Mode 2	17	0.85	5.0
Mode 3	19	0.95	5.3

Table 1: Explanation-based dialogue effectiveness.

3 Results

The baseline dialogue effectiveness results are reported in Table 1. As expected, injecting more information corresponds to better performances. However, it is to be highlighted that the model, initially unable to answer 0-shot, in our dialogical setting is able to answer correctly 9 of the 20 initial incorrectly answered questions. Moreover, we notice the larger accuracy rise from mode 1 to mode 2, indicating that providing the model with alternative options is particularly effective in guiding the student to the correct answer, results that are even outperformed when providing the model with human-written explanations. This confirms the need for carefully curated data in order to develop efficient explanatory dialogue systems, especially in the medical domain.

4 Conclusions

We presented an approach for developing synthetic explanatory dialogues for medical QA, highlighting the potential of dialogue-based explanations to develop and evaluate argument-based explanation models for medical QA systems. Baseline results suggest that dialogue-based explanations are a promising approach to improving the understandability of medical QA systems. In future work, we plan to move to open models, extend the approach to several languages, as well as analyze the arguments presented by both the Teacher and the Student to identify common argumentation strategies and their impact on the Student's understanding and ability to get to the correct answer.

[1] Code, data and example dialogues are provided at https://github.com/andreazaninello/MedExpDial

[2] With value = 0 if the answer is not among the options

References

Rodrigo Agerri, Iñigo Alonso, Aitziber Atutxa, Ander Berrondo, Ainara Estarrona, Iker García-Ferrero, Iakes Goenaga, Koldo Gojenola, Maite Oronoz, Igor Perez-Tejedor, German Rigau, and Anar Yeginbergenova. 2023. Hitz@antidote: Argumentation-driven explainable artificial intelligence for digital medicine. In *SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing.*

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Preprint*, arXiv:2404.05590.

Anonymous. 2024. Rewired: Instructional explanations in teacher-student dialogues. ACL ARR 2024 February Blind Submission.

Leona Cilar Budler, Lucija Gosak, and Gregor Stiglic. 2023. Review of artificial intelligence-based question-answering systems in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2):e1487.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R Pisani, and Kathryn Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Computers in biology and medicine*, 155:106649.

Gregory Kell, Angus Roberts, Serge Umansky, Linglong Qian, Davide Ferrari, Frank Soboczenski, Byron C Wallace, Nikhil Patel, and Iain J Marshall. 2024. Question answering systems for health professionals at the point of care—a systematic review. *Journal of the American Medical Informatics Association*, 31(4):1009–1024.

Valentin Li'evin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *Patterns*, 5.

Benjamin Molinet, Santiago Marro, Elena Cabrio, and Serena Villata. 2024. Explanatory argumentation in natural language for correct and incorrect medical diagnoses. *Journal of Biomedical Semantics*, 15(1):8.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Henning Wachsmuth and Milad Alshomary. 2022. "mama always had a way of explaining things so i could understand": A dialogue corpus for learning to construct explanations. *Preprint*, arXiv:2209.02508.

Niraj Yagnik, Jay Jhaveri, Vivek Sharma, Gabriel Pila, Asma Ben, and Jingbo Shang. 2024. Medlm: Exploring language models for medical question answering systems. *ArXiv*, abs/2401.11389.

Ackowledgements

We would like to thank the 2023/2024 students of the Language Technology Seminar at the University of Bologna for contributing to the annotation of this dataset. This work has been partially supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and by the ANTIDOTE project (CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002).