Reproducing MEDCOD Methodology
A Reproduction Study of a Medically-Accurate Dialog System

Anonymous submission

Introduction

The paper MEDCOD: A Medically-Accurate, Emotive, Diverse, and Controllable Dialog System (Compton et al. 2021) addresses the challenge of automating medical history-taking in a way that is both clinically accurate and engaging for patients. Traditional automated medical dialogue systems often lack flexibility, emotional intelligence, and contextual adaptability, making interactions feel robotic and impersonal. To bridge this gap, the authors propose MEDCOD, a hybrid system that integrates expert-driven rule-based decision-making with deep learning-based natural language generation (NLG). The system consists of three key components: (1) a Dialogue Manager that ensures medical relevance by selecting appropriate questions based on a structured knowledge base, (2) an Emotion Classifier that enhances patient engagement by predicting suitable emotional tones, and (3) an NLG module that generates diverse and natural-sounding questions. By balancing medical accuracy, empathetic interaction, and linguistic diversity, MEDCOD aims to reduce the documentation burden on physicians while improving patient experience and the efficiency of medical consultations.

Methodology

Specific Approach

The MEDCOD system was designed to address the challenge of generating medically accurate, emotionally aware, and diverse dialogues in the context of medical history-taking. The researchers took a hybrid approach, combining modular **expert-driven decision-making with deep learning-based natural language generation (NLG)**. This ensured that the system could follow structured medical knowledge while producing responses that sounded natural and empathetic.

At the core of MEDCOD is a **dialogue management system** that determines the next question based on a patient's responses. Instead of relying solely on machine learning, which might generate unpredictable outputs, the system leverages an **expert system** with a medical knowledge base. This expert system ensures that the dialogue remains medically consistent and follows a logical questioning pattern, similar to how a doctor would conduct a structured history-taking conversation. To make the responses feel more human-like, the system also integrates an **emotion classifier**. This classifier assesses the patient's emotional state and selects an appropriate emotional tone for the next response. By blending medical reasoning with emotion-aware dialogue generation, MEDCOD aims to create a more engaging and realistic interaction experience for patients.

To generate natural and diverse dialogue, MEDCOD incorporates **DialoGPT**, a version of GPT-2 fine-tuned for conversational tasks. However, instead of allowing DialoGPT to generate responses freely, the researchers introduced **control codes** that enforce medically relevant constraints while still allowing linguistic variation. These control codes ensure that the model can phrase a question in multiple ways while keeping the underlying medical meaning intact. Another key challenge in training the NLG module was the lack of large-scale medical dialogue datasets. To overcome this, the researchers used **GPT-3 for data augmentation**. GPT-3 was leveraged to generate additional training samples with variations in phrasing and emotional tone. This approach helped improve the robustness of the NLG model, ensuring it could handle a wide range of patient responses. The **emotion classifier** plays an important role in making MEDCOD's responses feel more natural. By analyzing the emotional tone of the patient's previous response, the classifier selects an appropriate emotional style for the next question. This prevents the system from sounding too robotic or impersonal, which is a common limitation in traditional rule-based medical chatbots.

To assess the effectiveness of the system, the researchers evaluated MEDCOD based on four key criteria: medical consistency, response diversity, emotional accuracy, and human-likeness.

Medical consistency was a crucial metric, as the system needed to ensure that its generated responses adhered to established medical knowledge and followed a structured decision-making process. The researchers manually checked whether the responses aligned with the expert system's expected questions and whether they avoided generating misleading or medically incorrect statements. **Response diversity** was measured by analyzing how varied the phrasing of the system's responses was while maintaining the same medical intent. This was evaluated using linguistic diversity metrics and manual review to confirm that the system wasn't repeating the same fixed templates but rather producing a range of natural-sounding variations. **Emotional accuracy** was assessed to ensure that the system correctly adapted its responses based on the emotion classifier's predictions. This was evaluated both through automated sentiment analysis and human reviewers who rated whether the system's responses matched the intended emotional tone. Finally, **human-likeness** was evaluated by comparing MEDCOD's responses to those of real doctors. Human judges rated the system's responses based on fluency, coherence, and relevance, determining how natural and engaging the dialogue felt compared to real human interactions.

Novelty, Relevance and Hypotheses to be Tested

Novelty and Relevance: MEDCOD system introduces a hybrid approach that integrates rule-based decision-making with deep learning, ensuring both medical accuracy and natural dialogue flexibility. Unlike traditional medical dialogue systems that either rely solely on rigid rule-based frameworks or fully data-driven models prone to inaccuracies, MEDCOD blends expert-driven structured knowledge with deep learning techniques. The system enhances linguistic diversity and emotional engagement, making it more human-like and effective for medical history-taking. By incorporating an emotion classifier and NLG module, the system adapts dynamically to patient responses, improving overall interaction quality.

Advantages Over Baselines: Compared to traditional medical dialogue systems, MEDCOD excels in three key areas: (1) Medical relevance ensured through its expert system-driven dialogue manager, which adheres to clinical protocols; (2) Empathy and patient engagement achieved via a supervised learning-based emotion classifier that predicts suitable emotional tones for responses; and (3) Natural language generation (NLG) using sequence-to-sequence deep learning models with attention mechanisms to generate diverse and human-like medical queries. These features help MEDCOD surpass template-based or rule-only systems, which often produce rigid, repetitive, or contextually inappropriate questions.

Key Hypotheses: The paper is built on two main hypotheses: (1) Enhanced patient engagement - A dialogue system that incorporates emotional and linguistic diversity will lead to greater user satisfaction and engagement than conventional rule-based approaches; and (2) Reduced physician workload - By automating medical history taking while maintaining accuracy and empathy, MEDCOD can alleviate administrative burdens on doctors, allowing them to focus on complex medical decision making. The evaluation results indicate that MEDCOD effectively balances medical accuracy, empathy, and linguistic richness, making it a more effective solution than traditional baseline models.

Ablations or Extensions Planned

The hypothesis of MEDCOD is legitimate and well-supported by their results. The hybrid approach effectively balances medical consistency, response diversity, and emotional awareness.

One key limitation is the restricted dataset access, which affects reproducibility and transparency, especially in healthcare research. Increased clarity on data collection and synthesis methods, along with a shift toward open-source medical knowledge bases, could improve accessibility and encourage wider adoption.

Another area for improvement is the paraphrasing module. Since MEDCOD was developed in 2021, it relies on GPT-3 (Davinci) with a basic system prompt for paraphrasing. Given advancements in LLMs like GPT-4, using more structured prompts and instruction-tuned models could significantly enhance paraphrase diversity and medical accuracy.

Additionally, the emotion classifier is currently limited to four broad categories (affirmative, empathy, apology, none). Extending it to include finer-grained emotional states such as concern, encouragement, or reassurance, and implementing multilevel classification could make the system more human-like and contextually adaptive in conversations.

Data Access and Implementation Details

To reproduce the computation presented in the MEDCOD paper, we need access to both the dataset and the model used in their implementation. The authors have made the code repository publicly available at https://github.com/curai/curai-research/tree/main/MEDCOD, but pre-trained model weights are not provided. Therefore, we will need to retrain the model following their methodology.

MEDCOD is trained using multiple datasets, primarily derived from a proprietary medical knowledge base (KB) and augmented data. These include a doctor-edited dataset for emotion classification, simulated clinical cases, GPT-3-generated paraphrased questions, and dialogue samples labeled with emotion types. Additionally, a combined medical conversations dataset is used to train the natural language generation (NLG) model. Since these datasets are proprietary, direct replication is challenging. To address this, we will generate synthetic data for each component using small samples available in the codebase and leverage large language models (LLMs) to create a sufficiently large dataset for training.

Originally, with access to the full dataset and model artifacts, reproducing the computation would have required significant computational resources. However, since we are working with a smaller synthetic dataset, our initial iterations suggest that the process is feasible. We will utilize free GPU resources available on Google Colab to support training and experimentation.

Overall, while exact reproduction of MEDCOD is challenging due to dataset constraints, we can implement a functionally equivalent system with careful dataset engineering and computational optimizations. We will use the existing codebase as a foundation and modify it to accommodate our synthetic dataset.

References

Compton, R.; Valmianski, I.; Deng, L.; Huang, C.; Katariya, N.; Amatriain, X.; and Kannan, A. 2021. MEDCOD: A Medically-Accurate, Emotive, Diverse, and Controllable Dialog System. *arXiv preprint arXiv:2111.09381*.