Improving efficiency, reliability of AI medical summarization tools

February 22 2024, by Mary Fetzer

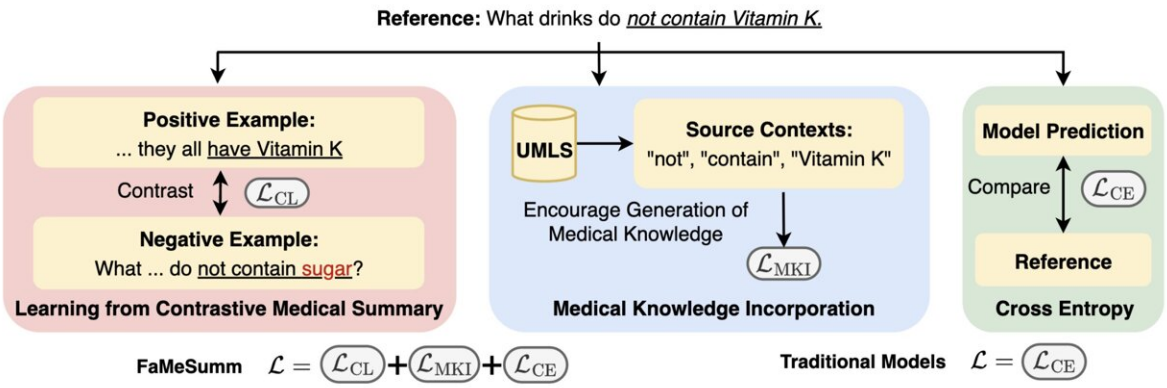

Diagram of FAMESUMM architecture with an example reference summary. The underlined part in the reference contains a medical term ("Vitamin K") and its context ("do not contain") that are modeled by FAMESUMM. Credit: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023). DOI: 10.18653/v1/2023.emnlp-main.673

Medical summarization, a process that uses artificial intelligence (AI) to condense complex patient information, is currently used in health care settings for tasks such as creating electronic health records and simplifying medical text for insurance claims processing. While the practice is intended to create efficiencies, it can be labor-intensive, according to Penn State researchers, who created a new method to streamline the way AI creates these summaries, efficiently producing

more reliable results.

In their work, which was published as part of the *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* in Singapore last December, the researchers introduced a framework to fine-tune the training of natural language processing (NLP) models that are used to create medical summaries.

"There is a faithfulness issue with the current NLP tools and machine learning algorithms used in medical summarization," said Nan Zhang, a graduate student pursuing a doctorate in informatics at the College of Information Sciences and Technology (IST) and the first author of the paper. "To ensure records of doctor-patient interactions are reliable, a medical summarization model should remain 100% consistent with the reports and conversations they are documenting."

Existing medical text summarization tools involve human supervision to prevent the generation of unreliable summaries that could lead to serious health care risks, according to Zhang. This "unfaithfulness" has been understudied despite its importance for ensuring safety and efficiency in health care reporting.

The researchers began by examining three datasets—online health question summarization, radiology report summarization, and medical dialogue summarization—generated by existing AI models. They randomly selected between 100 and 200 summaries from each dataset and manually compared them to the doctors' original medical reports or source text, from which they were condensed. Summaries that did not accurately reflect the source text were placed into error categories.

"There are various types of errors that can occur with models that generate text," Zhang said. "The model may miss a medical term or change it to something else. Summarization that is untrue or not

consistent with source inputs can potentially cause harm to a patient."

The data analysis revealed instances of summarization that were contradictory to the source text. For example, a doctor prescribed a medication to be taken three times a day, but the summary reported that the patient should not take said medication. The datasets also included what Zhang called "hallucinations," resulting in summaries that contained extraneous information not supported by the source text.

The researchers set out to mitigate the unfaithfulness problem with their Faithfulness for Medical Summarization (FaMeSumm) framework. They began by using simple problem-solving techniques to construct sets of contrastive summaries—a set of faithful, error-free summaries and a set of unfaithful summaries containing errors.

They also identified medical terms through external knowledge graphs or human annotations. Then, they fine-tuned existing pre-trained language models to the categorized data, modified objective functions to learn from the contrastive summaries and medical terms, and made sure the models were trained to address each type of error instead of just mimicking specific words.

"Medical summarization models are trained to pay more attention to medical terms," Zhang said. "But it's important that those medical terms be summarized precisely as intended, which means including non-medical words like no, not, or none. We don't want the model to make modifications near or around those words, or the error is likely to be higher."

FaMeSumm effectively and accurately summarized information from different kinds of training data. For example, if the provided training data comprised doctor notes, then the trained AI product was suited to generate summaries that facilitate doctors' understanding of their notes.

If the training data contained complex questions from patients, the trained AI product generated summaries that helped both patients and doctors understand the questions.

"Our method works on various kinds of datasets involving medical terms and for the mainstream, pre-trained language models we tested," Zhang said. "It delivered a consistent improvement in faithfulness, which was confirmed by the medical doctors who checked our work."

Fine-tuning large language models (LLMs) can be expensive and unnecessary, according to Zhang, so the experiments were conducted on five smaller mainstream language models.

"We did compare one of our fine-tuned models against GPT-3, which is an example of a large language model," he said. "We found that our model reached significantly better performance in terms of faithfulness and showed the strong capability of our method, which is promising for its use on LLMs."

This work contributes to the future of automated medical summarization, according to Zhang.

"Maybe, in the near future, AI will be trained to generate medical summaries as templates," he said. "Doctors could simply doublecheck the output and make minor edits, which could significantly reduce the amount of time it takes to create the summaries."

More information: Nan Zhang et al, FaMeSumm: Investigating and Improving Faithfulness of Medical Summarization, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023). DOI: 10.18653/v1/2023.emnlp-main.673

Provided by Pennsylvania State University