innovate
GenAI Use Case

phactMI

Leveraging AI to Generate Medical Responses

Background:

Medical Information personnel spend a significant amount of time researching and understanding authoritative, evidence-based information that is used to answer healthcare professionals' questions. They must condense and capture information into concise and easily digestible summaries used to respond to specific HCP inquiries. Our challenge was to pilot generative Artificial Intelligence (gen AI) technology that could summarize existing publications and generate a scientific response document (SRD).

Our early work used transformer models but didn't provide the best model for extracting content from publications. We then partnered with a technology company to use several text generation and summarization options including GPT-3 (generative pre-training transformer). The other models and GPT-3 both did a competent job summarizing, however GPT-3's ability to paraphrase without plagiarizing led us to use GPT-3 in our proof of concept (POC).

Methodology:

1. We started by establishing the components of a "good" publication summary used in SRDs. We wanted to understand and "classify" the decisions being made by humans when summarizing clinical trials and then replicate those classifiers (patient population, study design, objectives, dosage and administration, endpoints, efficacy, and adverse events) in training models for the tool.
2. Next, we aligned on the content generation approach: upload applicable publications; extract content (machine learning); summarize and paraphrase the content (transformer), and filter unusable content (machine learning).
3. Finally, we applied our "authoring/content generation" learning to a POC application in medical information –the SRD Assembler.

Results:

Topic relevant publications were uploaded into the POC content generator/SRD assembler. We used the POC tool to perform the content extraction and to assign an associated relevance score. The technology generated summaries for each of the uploaded publications. Then, we conducted a review of the content summarization by comparing text extracted verbatim from the publication to AI-generated summarized text and AI-generated paraphrased text. As we reviewed the results from the initial stage,

we noticed that some of the classifiers were able to extract relevant text more reliably than others. The most reliable classifiers were those extracting treatment and safety related information due to treatment containing recognized unique words such as product names and adverse events using common and consistent names for medical conditions.

In an iterative process, single summaries were selected, and the extracted content used to generate the summary was reviewed to determine importance and relevance to the summary. In doing these comparisons, we calculated two values between the original extracted text and its counterpart summarized/paraphrased text. These were semantic similarity and text similarity. We initially evaluated our ability to produce summarized content that was saying the same thing (high semantic similarity) but not plagiarizing (low text similarity). These results were very strong in that we were able to maintain high semantic similarity without having to mimic the exact text. Extractions were removed if irrelevant, and summaries were regenerated as needed.

Following the extraction and summarization process, the final and most important assessment of the generated content was the "usefulness judgement" to ensure the content was something that a medical author would use in their SRD. The initial feedback from our first SRD author was promising as we were able to effectively develop a filter to remove non-useful content and retain useful content with > 80% accuracy. However, as more writers' feedback was obtained, the predictions were less accurate dipping below 50%.

POC Conclusions and Applications:

- GPT models can successfully generate publication summaries that can be used in SRDs.
- Filtering logic to reduce the number of unusable responses needs to be built in.
- Domain-focused GPT models may be able to extract content as well or better than traditional machine learning approaches.
- To improve accuracy of determining useful content more analysis is needed to determine what factors the author is using to pull content into their SRDs. Our classifier themes were the most obvious, however there are other decisions that could possibly warrant using additional classifiers in tandem with the ones already developed.
- A significant drawback of current GPT models is the lack of score or confidence in a prediction.
- Tracing back to the exact lineage with GPT is much more difficult than with machine learning. Until this evolves, QA of the generated responses is time consuming and tedious.
- Companies need to use some form of GPT/machine learning to generate usefulness data that drive better outcomes in content selection.
- Content generation using Gen AI and structured authoring are symbiotic and not competing.
- GPT ultimately is the future as training traditional machine learning models requires substantial up-front effort.
- Results support a "Reuse before Generate, Generate before Write" philosophy.

For more information visit phactMI.org.