AMERICAN MEDICAL WRITERS ASSOCIATION AMWA JOURNAL V38 N1 / 2023

ARTICLE

Leveraging Artificial Intelligence, Natural Language Processing, and Natural Language Generation in Medical Writing

Deepak Palasamudram[1]; Karun S. Karunakaran[2]; Prakhar Gaur[3]; Akshatha Miyal Kamath[2]; Pramit Saha[4]; Tina Purushotam[5]/
[1]Associate Vice President, Architecture and Design Group, Healthcare, Insurance and Life Sciences, Infosys Limited, India; [2]Architect, Architecture and Design Group, Healthcare, Insurance and Life Sciences, Infosys Limited, India; [3]Consultant, Life Sciences Domain Consulting Group, Infosys Limited, India; [4]Project Manager, Architecture and Design Group, Healthcare, Insurance and Life Sciences, Infosys Limited, India; [5]Digital Specialist Engineer, Architecture and Design Group, Healthcare, Insurance and Life Sciences, Infosys Limited, India

ABSTRACT

Medical writing is a process that generates a variety of documents in the biomedical domain, including but not limited to clinical reports, regulatory reports, protocol documents, patient narratives, plain language summaries, and so on.[1] Medical writing is complex and time-consuming because a writer must refer to multiple sources, sift through a large volume of documents, maintain data integrity, perform review of literature, do interpretation of results, summarize, and so on.

These challenges can be addressed and minimized substantially by adopting artificial intelligence, specifically cognitive search, natural language processing (NLP), and natural language generation (NLG) models and other techniques. Given the recent advances in language models for NLG, the time is ripe for a product in the medical writing domain that integrates and automates search capabilities, provides cognitive processing, and generates content using NLG.

This white paper takes scientific manuscript writing as an example to provide insights into the way NLP and NLG can augment, automate, and expedite the process of writing a wide variety of biomedical documents. It looks at the current limitations of technology and ways to address those. Finally, it provides recommendations on how these technologies can be used to create a single system or product. Such an approach has the potential to expand into multiple areas in the biomedical domain, with medical writing as the first challenge.

GLOSSARY

Natural Language Processing (NLP): The branch of artificial intelligence (AI) that enables computers to process human language and understand the meaning, intent, and sentiment of the text, much like a human being can.

Natural Language Generation (NLG): The branch of AI that enables computers to produce human language that approximates content generated by a human being.

Recommendation Model: A system that uses machine learning to predict content that is relevant for a user in a given context. The predictions are often combined with a ranking system that enables users to see the most relevant recommendations first.

Named Entity Recognition (NER): A process by which text is classified into predefined categories like drug name, disease name, location. Also, depending on the context, it can differentiate between "apple" (fruit) and "Apple" (corporation).

Large Language Model (LLM): Large Language Models (LLMs) are artificial intelligence tools that can read, summarize and translate texts and predict future words in a sentence letting them generate sentences similar to how humans talk and write.[11] Eg., GPT-3, GPT-J, BART, BERT, t5.

Natural Language Query Understanding (NLQU): This is a capability of the search system to understand a search query written in natural language. This is achieved by LLM based search systems. Eg. "What is the second largest land animal in the world?"

INTRODUCTION

CURRENT MEDICAL WRITING MARKET

According to Grand View Research, the global medical writing market size that was valued at US $3.4 billion in 2019 is expected to expand to US $7.77 billion by 2027 at a compound annual growth rate of 10.9%.[2] The cost spent on content generation continues to rise.

EXISTING MEDICAL WRITING PROCESS

Medical writing is a complex and manually intensive process. The process of medical writing involves the following steps

a) Understanding the content brief
b) Review of literature
c) Collation of the results, methods, and discussion sections

ISSN 2163-5315

d) Authoring the manuscript and maintenance of data integrity in the process
e) Reviewing the authored content
f) Copy editing
g) Approval and sign off
h) Electronic publishing

Medical writers spend 2 to 3 weeks researching across multiple data sources and a large corpus of documents (nearly 1 million new articles are added yearly to just PubMed).[3]

The review of literature is the most time-consuming step in the medical writing process. This requires a domain expert to first search for and then read through the text of the articles on a particular subject area. The goal of this step is to synthesize the existing knowledge in a particular subject area. In the case of writing a scientific manuscript for a clinical trial, the review of literature must cover several subtopics in the therapy area of concern. All the subtopics require individualized search strategies irrespective of the therapeutic area. Medical writers use several literature databases like PubMed, Scopus, Ovid, and Cochrane. This also introduces the risk of missing out on relevant literature, making this task not only time-consuming, but also error-prone. The aforementioned tasks require multiple individuals to complete it in a reasonable amount of time, each one concentrating on a particular subset of the overall document.

The final challenge lies with the summarization step, when information gleaned from several published articles is summarized. The risks here are missing the important points as well as accidentally not including a relevant reference.

In writing the results section of a manuscript, data may need to be collated from a source document like a clinical study report (CSR) and adding it to the manuscript in a particular format. This can involve aggregating and summarizing the data, creating plots, or writing a narrative for a particular set of data. A great example are the tables for adverse events. This step may introduce quality issues if not done carefully. Obviously, the power of using a computer to automate data analytics is well known, and natural language generation (NLG) provides tools to create narratives summarizing tabular data accurately.

Products with authoring workflows that allow collaboration on a single document by multiple people have been in use for more than a decade.[4] In addition to these, functionalities like referencing and text formatting according to journal requirements have also been in use. These capabilities can come from various tools and techniques, which would require integration of many products or tools into a single system.

ROLE OF ARTIFICIAL INTELLIGENCE IN MEDICAL WRITING

A system that can automate and assist with these tasks would help mitigate many of the challenges and risks described before. Artificial intelligence (AI) has several interesting possibilities for transforming any industry. Nowhere are its applications more relevant than for life sciences and pharmaceutical, regulatory, and medical writing for creating documents such as scientific manuscripts, fact sheets, literature reviews, disease awareness, and oral posters.

A canonical use case for the application of AI is the process of the review of literature that is done as part of research work. In the review of literature process, the researcher is required to use their language and domain knowledge to summarize the various published articles on a topic. This is done to summarize the state of the art in the field. NLG models are being used to automate this step; at the same time, manual intervention is required before the machine-generated text can be submitted for publication.

Many functionalities are required to automate the medical writing process that are elaborated on in the following sections.

Cognitive Search for Literature Survey and Recommendation

The review of literature requires multiple subtopics in a disease/therapeutic area to be comprehensively covered. This requires individualized search strategies for each subtopic, with the search itself carried out across multiple databases. The next task is to read through the top hits for each subtopic to identify the relevant content in that published article.

AI or machine learning (ML) can help in automating the literature search, content extraction, content enrichment (which includes named entity recognition [NER]), and intent detection in the context of life sciences and pharmaceuticals. This enables advanced unified searching across multiple data sources and databases.

Secondly, once trained, the recommendation models are used to identify the relevant sentences or paragraphs in the articles. Automation of this step can enhance efficiency of the most time-consuming component in the literature review process. In addition to automation, the system can perform citation management, an essential part of any medical or technical document.

Narrative Generation

The generation of narratives from structured data by applying NLG has been in the life sciences domain for more than a decade. Previously, it was done using hardcoded text as part of code, and now, NLG models can generate text up front for the structured data that is being processed. The text or narrative generated by the NLG model can, depending on the models used, involve ML algorithms (deep learning) or preset parameters. The output from either of these or a combination of both is a narrative as would have been written by a medical writer.[5]

Summarization

The most recent developments in the NLG space have enabled models to summarize large texts. The initial models were trained and built using news articles because they provide human-generated text and summaries. Now, models are being trained on the published medical corpus available as peer-reviewed scientific articles. Further refinement of the models specific to diseases and drugs are needed in automating medical writing.

There are 2 types of summarization techniques possible using NLG models: extractive and abstractive.

Extractive summarization involves identifying important subsets of sentences from the original text in toto and forming a summary comprising such sentences. This type of summarization is useful if the author decides to select multiple sources from the recommendations and to rewrite the text on their own after the summary gets generated.

Abstractive summarization reproduces important material in a new way after examination and interprets the text using NLG capabilities, simulating how humans do a review of literature.[6,7]

Abstractive summarization mimics how an author would write a synthesis of existing literature in their own words along with the references used. Examples of such models include GPT-3, t5, BERTs, and BARTs.[8] Abstractive summary is useful when the authors want an abstract of the selected recommendations. This kind of summary, along with reference metadata, addresses issues related to plagiarism because this is not an exact reproduction of text from the sources but a generation of original text.

Figure 1 depicts one such example of extractive and abstractive summarizations.

KEY CHALLENGES IN APPLYING AI TO MEDICAL WRITING

The key challenges to applying AI in medical writing include

- Ingesting documents from diverse sources and variety of formats. Apart from a pharmaceutical company's internal data sources, there are multiple external sources like PubMed articles, regulatory documents, clinical trial documents, protocols, clinical study reports, and press releases. This necessitates dealing with different document formats and structures like native PDFs, Docx, XML, HTML, and scanned documents.
- Understanding document semantics and content, extracting key entities unambiguously, capturing synonyms based on scientific ontologies, identifying contexts and intents from medical content in the context of life sciences and pharmaceuticals. This requires compositional semantic analysis that includes word sense disambiguation and relationship extraction that is relevant in biomedical literature.
- Reranking cognitive search results for better search relevance. This requires adoption of learning to rank

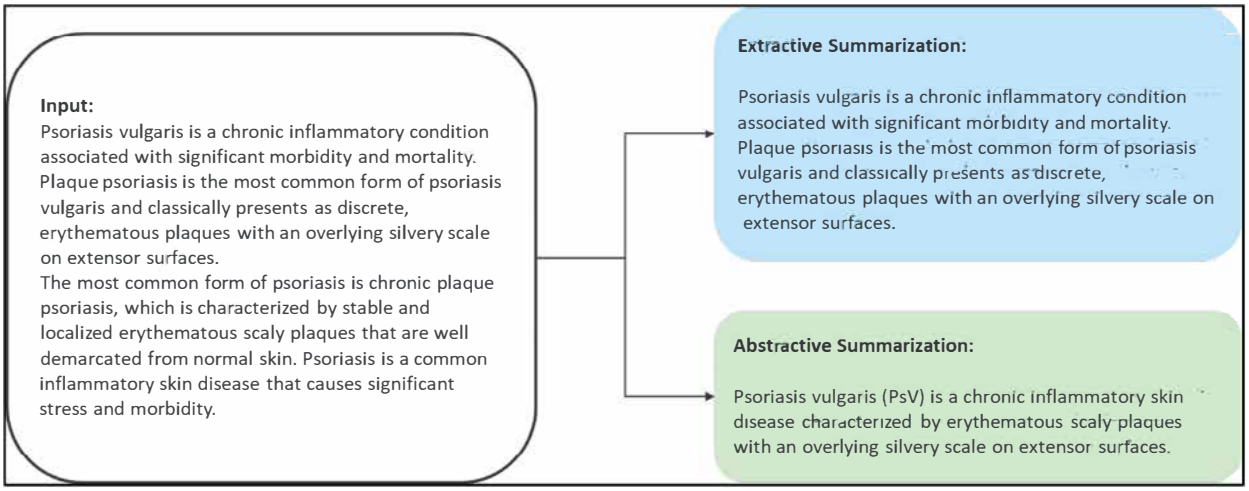

Figure 1. Extractive and abstractive summaries from given input text.

also know as machine-learned ranking. This process re-ranks results from search engines in the medical writing context for content such as the mechanism of action for drugs or disease epidemiology, etc.[9]
- Combining multiple modes of intelligence such as natural language processing (NLP), NLG, deep learning, language models, lexicons, and ontologies into a state-of-the-art AI-based platform for medical writing.
- Removing biases from algorithms. Potential biases can creep in at various steps, from the curation of training datasets, to feature engineering, model choice, and implementation. Detecting and removing algorithmic biases will entail evaluating it via a thorough understanding of the algorithm's role and the context in which it is deployed.

BRINGING IT ALL TOGETHER

As discussed earlier, one of the objectives of this article is to define the architecture and components of a system or product that will automate the process of medical writing significantly. Such a system should have the end-to-end ability to ingest documents, identify the entities in those documents, provide them as search results based on user queries, and generate summaries based on user-selected documents. These features require adoption of the various AI techniques discussed previously.

AI horizons have seen a strategic shift from conventional ML (with a focus on augmenting intelligence) to deep learning (enabling higher accuracy and predictability), and now to the responsible, transparent generative AI.

The key emerging trends for language processing and generation include

- Adoption of deep learning and transfer learning architectures driving accuracy, performance, and speed.
- The NLP shifts from extraction of isolated entities to abstractive reasoning and language models.
- Using models for text critiquing, information retrieval, question answering, summarization, gaming, text generation, and translation.

With state-of-the-art pretrained language models (eg, GPT-3, GPT-J, BART, BERT) that can be fine-tuned for the biomedical domain, the system can generate human-like summarizations and narratives.[10] Consequently, the text summarization exercise and the final document generation can be reduced to a few days rather than a few weeks, even after accounting for the final manual review and approval processes.

Moreover, with all workflows automated, the scope of error is minimized, contrasted with the current manual process (Figure 2).

To bring about these efficiencies, the AI-led platform for life sciences is envisaged to encompass the following key features:

- Unified search across multiple internal and external databases
- Built in deep learning models for article recommendation in the context of pharmaceutical clinical trials, regulatory intelligence, and medical research
- State-of-the-art language models fine-tuned for life sciences for text summarization and NLG tasks
- Real-time data ingestion of structured or unstructured documents from varied data sources (scientific articles from PubMed, regulatory sources like the US Food and Drug Administration (FDA), European Medicines Agency, CSRs, and protocol documents, etc)
- NLP-based automatic document structure extraction, content enrichment, and sentiment analysis
- Dynamic document editing features leveraging scientific lexicons and ontologies
- Workflows for collaborative medical authoring
- Content citations (ability to refer to original sources from a machine-generated summary)
- Templatization of the final document based on the need

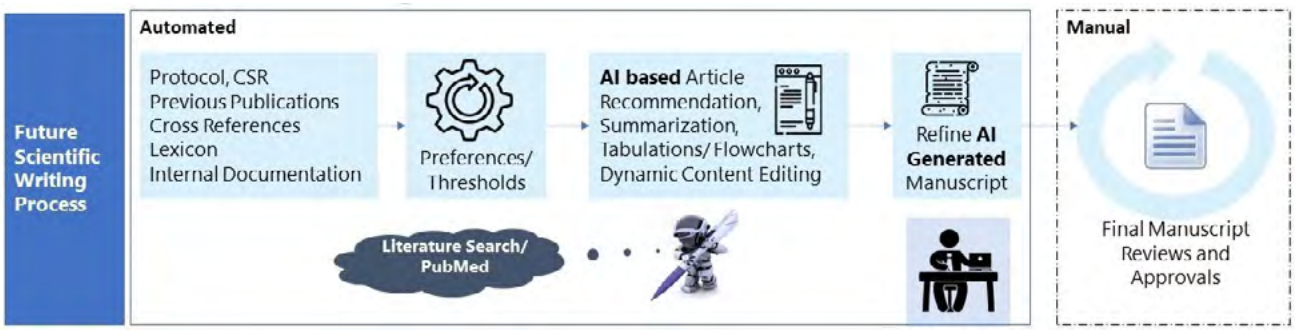

Figure 2. Automated process of medical writing. AI, artificial intelligence; CSR, clinical study report.

Medical writing typically involves authoring contents like medical manuscripts, posters, or clinical study reports with predefined templates for each content type. An authoring template is not just a bare-bone skeleton for content authoring but a composite of individual sections, the onboarding of which entails data ingestion, article recommendation, and content summarization steps (Figure 3).

Let us illustrate this through an example of the introduction section of a typical manuscript. This section includes content primarily from PubMed articles contextualized for disease description, epidemiology, burden of disease, and a drug mechanism of action.

The following activities are required for the generation of an introduction section of the manuscript.

- For data ingestion, PubMed articles are considered, and indexing is configured for relevant article sections like "Abstract" and "Introduction." NLP pipelines are used for the classification of sentences as belonging to categories like "description of disease," "burden of disease," "disease epidemiology," and "mechanism of action." These create labels and do NER for diseases, drugs, molecules, and so on.
- For article recommendation, natural language query understanding pipelines for intents like classification contexts are defined. Search ranking rules and boosting criteria are refined as required.
- For content summarization, based on the section specific summarization or narrative needs, the platform evaluates the available language models. For configuration initial training, samples are curated for platform-suggested language model fine-tuning, and pipelines are defined for subsequent active learning.

Figure 4 provides a schematic view of the platform architecture. The document sources will not only be external in nature like PubMed, Ovid, and ClinicalTrials.gov, but also

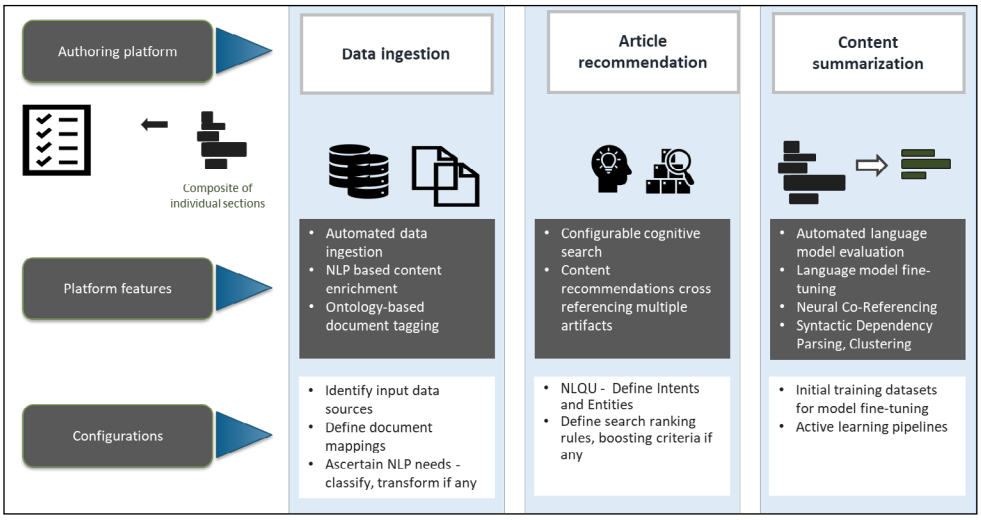

Figure 3. Section onboarding in medical writing platform. NLP, natural language processing.

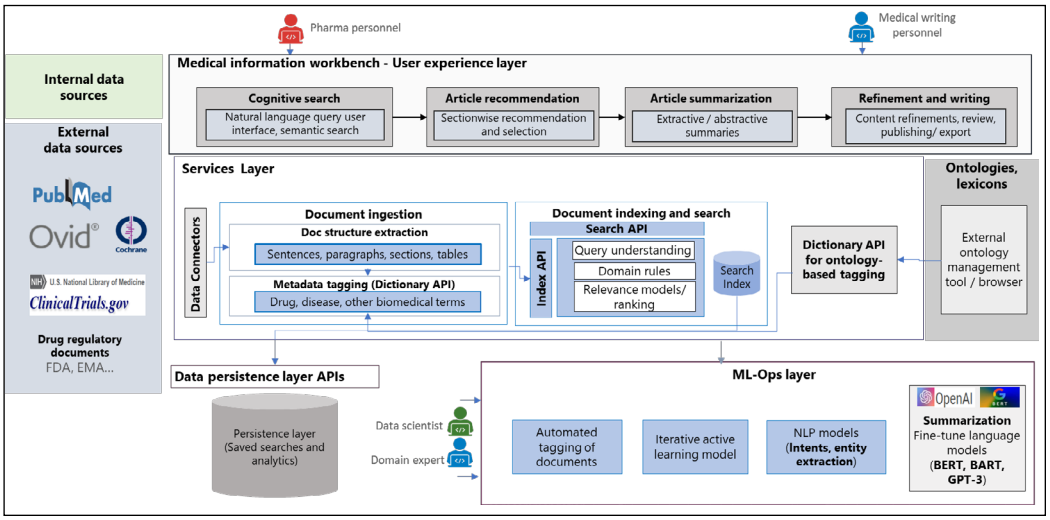

Figure 4. A schematic view of the platform architecture. API, application programming interface; EMA, European Medicines Agency; FDA, United States Food and Drug Administration; ML, machine learning; NLP, natural language processing.

regulatory documents like those published by the FDA and document sources that are internal to any organization deploying the platform.

During the ingestion of the documents, document structure is extracted and tagged with metadata that helps in indexing and classifying the content for future use. Such extraction includes classification of sentences and paragraphs and sections as dealing with different drugs, diseases, and other biomedical terms.

When a user searches for documents, the questions posed by the user in plain English are translated into machine-readable queries that are then searched against the indexed documents. The results are then ranked according to the rules and boosting criteria used and returned to the user as recommendations.

Once the user selects the documents identified for summarization, NLG is used to generate extractive or abstractive summaries of the selection.

CONCLUSION

AI and ML, combined with NLP and NLG, promises to benefit the medical writing process by reducing the manual aspects of the work by automating many steps, in addition to improving quality and reliability. The time and effort thus saved can be substantial to large organizations that often spend a considerable amount of both during the lifecycle of a drug.

Author declaration and disclosures: *The authors acknowledge the team's learning from building the Cognitive Search and Medical Writing Platform at Infosys.*

Author contact: *DEEPAKPN@infosys.com*

References

1. Alexander L, De Milto L, Kryder C. Ultimate guide to becoming a medical writer. AMWA. Accessed September 2022. https://info.amwa.org/ultimate-guide-to-becoming-a-medical-writer
2. Medical writing market size, share & trends analysis report by type (clinical, regulatory), by application (medical journalism, medico marketing), by end use, by region, and segment forecasts, 2022-2030. Grand View Research. Published April 2022. Accessed September 2022. https://www.grandviewresearch.com/industry-analysis/medical-writing-market
3. PubMed production statistics. Accessed May 2022. https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html/
4. Google Docs. Wikipedia. Accessed September 2022. https://en.wikipedia.org/wiki/Google_Docs#History
5. Cawsey AJ, Webber BL, Jones RB. Natural language generation in health care. *J Am Med Inform Assoc.* 1997;4(6):473-482.
6. Teo L. Report is too long to read? Use NLP to create a summary. Towards Data Science. Published October 29, 2020. Accessed May 2022. https://towardsdatascience.com/report-is-too-long-to-read-use-nlp-to-create-a-summary-6f5f7801d355
7. Kan MY, McKeown KR, Klavans JL. Applying natural language generation to indicative summarization. arXiv. Preprint posted online July 16, 2001. doi:10.48550/arXiv.cs/0107019
8. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Paper presented at: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA. Accessed September 2022. doi:10.5555/3295222.3295349
9. Wang R, Shivanna R, Cheng DZ, et al. DCN V2: improved deep & cross network and practical lessons for web-scale learning to rank systems. arXiv. Preprint posted online October 20, 2020. doi:10.1145/3442381.3450078
10. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. The University of British Columbia. 2018. Accessed September 2022. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf
11. Parthasarathy discusses implications of Large Language Models. University of Michigan. Accessed January 2023. https://fordschool.umich.edu/news/2022/parthasarathy-discusses-implications-large-language-models