Almirall | Spain | Innovation & AI

Scientific Paper Search Engine

Client profile

In healthcare's ongoing transformation, the integration of generative AI represents a significant advancement. NTT DATA and Microsoft have collaborated to develop platforms that harness the potential of Azure OpenAI Service in real-world applications. By automating information extraction from drug-related papers, this innovation offers researchers an efficient tool, reducing time and costs in analyzing scientific advancements.

Why NTT DATA?

- We build innovative, industry-leading solutions that grow enterprises' revenue and keep them ahead the competition.
- We take advantage of the growing convergence of IT and connectivity services to connect people and things.
- We manage companies' applications, data and infrastructure to decrease costs and create greater efficiencies.

> "It's no news that AI represents a paradigm shift in the market right now, and we see generative AI models from Microsoft Azure Open AI Service' will help companies in any industry and across a wide range of use cases."

David Pereira, Head of Data & Intelligence – NTT DATA

Business need

Researchers spend countless hours manually reviewing drug-related papers and extracting relevant information. Automating the information extraction process and enabling easy access to papers with specific characteristics would significantly decrease the time and cost involved in analyzing scientific advancements. This would provide researchers with a more efficient tool for their work.

Solution

NTT Data devised a robust solution for RWE medical papers, employing a three-part pipeline. The process involved OCR for text and table extraction in preprocessing, utilizing section titles to segment papers and remove unnecessary data. Information extraction indexed paragraphs and tables with metadata, employing open-source language models for document retrieval and a Q&A system with GPPT-3 for generating answers. Excel generation used a JSON configuration file to produce output files, saving responses or color-coded cells denoting concept/score presence. This comprehensive approach ensured the efficient processing and extraction of valuable information from real-world evidence in medical papers.

Outcome

- **Increased Efficiency**: The system's high accuracy (88%) and ability to answer 29 concepts from RWE medical papers streamline information extraction, saving time and effort. This efficient access to critical insights accelerates research and analysis processes, benefiting researchers by saving valuable time and resources
- **Error Reduction**: The system surpasses human extraction in over 50 instances, reducing errors for higher data quality. This improves analysis and decision-making accuracy.
- **Scalability and Versatility**: Researchers can include a high volume of papers, expanding their analysis scope for comprehensive insights
- **Adaptability and Future Potential**: The system's adaptability allows for refining questions and considering specific parameters, enabling expansion into new research areas.

TECHNICAL SPECIFICATIONS

An easily accessible knowledge database has been established by generating an Excel file with relevant answers and a Web App for self-searching within RWE medical papers. This was done by creating a robust paper preprocessing pipeline, indexing data, and implementing a Q&A system. As a result, the search engine can efficiently retrieve and provide precise responses from the indexed information.

Indexing pipeline: The newly preprocessed paragraphs and tables, along with associated metadata, are indexed in Elastic (ref). For document retrieval, two open-source models are employed: the BM25 model for keyword-based retrieval and the DRP model for semantic sentence meaning.

Q&A system: After posing a question, the system retrieves documents and sends the top 3 answers, along with the question, to the Q&A system. It constructs a customized prompt for the question, focusing on obtaining tables or paragraphs, and sends it to a GPT-3 via the DaVinci-003 model API. The resulting response is then processed and returned to the user.

Document creation: To produce an Excel output file, the system reviews the JSON configuration file containing the required questions. It retrieves pertinent documents, generates prompts, forwards them to GPT-3, and records answers in the file. This yields the final Excel document.

Web App: The Web App features a user-friendly back-office page for uploads and data extraction to download the outputted Excel files, as well as a search engine so users can ask questions to about the uploaded documents with filters for document type and result limits

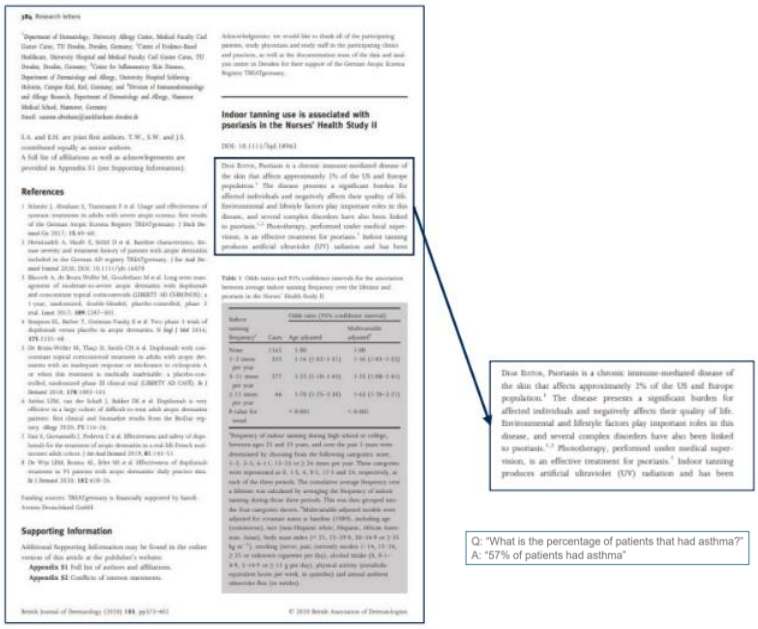

Fig 1. Information extraction pipeline with question classification and prompt engineering. Uses a search flow structure.

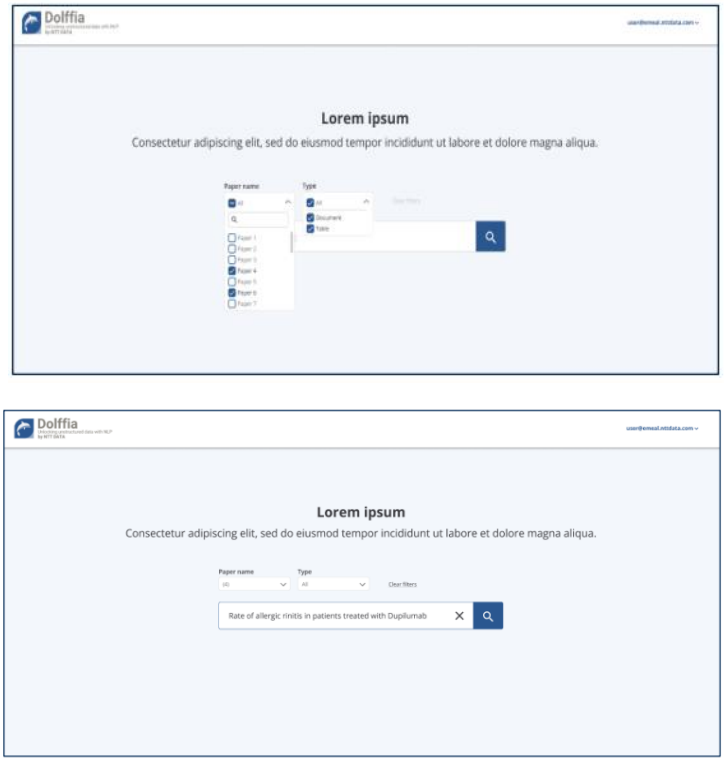

Fig. 2 Front –Home page