Extracting Key Concepts and Insights from Scientific Papers using NLP

Degree programme: BSc in Computer Science | Specialisation: Data Engineering
Thesis advisor: Prof. Dr. Erik Graf
Expert: Reto Trinkler

This thesis presents an approach to improve scientific literature searching using ML and NLP techniques. NER and summarisation algorithms enable efficient facet search and rapid assessment of articles, contributing to the advancement of research and fostering collaboration and innovation.

Introduction and Motivation:

The number of scientific papers published is increasing at an ever faster rate. As a result, knowledge and innovation are being created and progress is being made. However, the resulting information overload can make it difficult for researchers and students to effectively identify, access and analyse relevant information. This challenge is even more pronounced in areas such as AI and NLP, where researchers must keep up with the rapid advances in these and several related research areas. To overcome these hurdles, this project explores an approach that uses ML and NLP techniques to improve the process of searching and presenting scientific literature. The aim is to understand the challenges of scientific literature retrieval, identify promising applications for ML and NLP, and develop a user-friendly and intelligent system capable of analysing scientific texts.

Project objectives:

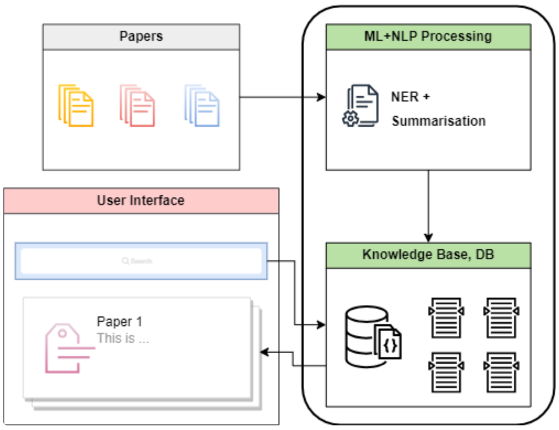
The objectives of the project include the exploration of the landscape for searching scientific information, the development of a state-of-the-art solution to improve the search and retrieval process, the design of a prototype, the comparison of the solution, and the identification of future research and development directions. Ideally, such a system would improve researchers‘ interaction with the scientific literature, thereby accelerating scientific progress, fostering collaboration and promoting innovation.

Enhancing Search with ML+NLP:

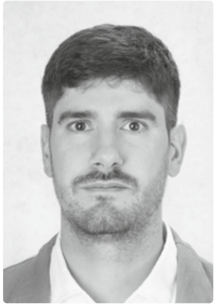
The integration of machine learning and natural language processing technologies into scientific literature search systems can significantly improve the search experience and efficiency for researchers. Named Entity Recognition (NER) provides a method for automating the recognition of defined categories within a text. It can help to extract key entities such as names of researchers, institutes, key concepts, methodologies or even specific research datasets. By enabling faceted searching through NER, researchers and students can refine their search results beyond keywords. Similarly, the use of automatic summarisation algorithms greatly enhances the utility of a search system. An article summary provides a condensed view of the content of the article, allowing users to quickly assess the relevance of the article to their query.

Findings:

The analysis of the use cases of language processing technologies in the field of scientific literature retrieval has shown that two ML+NLP techniques suitable for search systems are Named Entity Recognition (NER) and summarisation algorithms. These can be used to automatically generate additional information about publications, to provide the user with an improved navigation dimension in the result view, and to present the content of the paper in a concise manner. The solution design overview shows how the different components can interact in such a system. Full text and metadata of scientific publications are fed into an NLP pipeline. The articles enriched with NLP results are stored so that a UI prototype can query these results to show how researchers can benefit from the enriched information when searching the literature. Since the release of ChatGPT by OpenAI, many vendors have launched such systems.



Components of the ML+NLP based literature search system



Nicolas Scheurer