Karel 3.0

Karlo Šerbetar, Elguja Archemashvili, Varvara Prokacheva, Martin Ondra, Leonid Burbygin, Tomáš Kubeš

Bene Meat Technologies a. s.

Introduction

Innumerable availability of scientific literature, combined with the lack of systematization, poses a challenge for scientists in finding the necessary information. Often, reading an article is required to determine that it is not relevant. For example, it could be a difference in an experimental design, research methods, or an ultimate goal. This leads to the loss of valuable time for research and experiments.

Recently developed large language models (LLMs) can be utilized to process the scientific library in a semantic and systematic way. However, as the models are inherently probabilistic, they can generate factually incorrect information (hallucinate). Thus they are unsuitable for information retrieval on their own.

Therefore our team has developed the Scientific Paper Analyzer aka Karel 3.0 designed to streamline the research process by combining several NLP technologies to filter out irrelevant articles and provide in-depth summary of pertinent content.

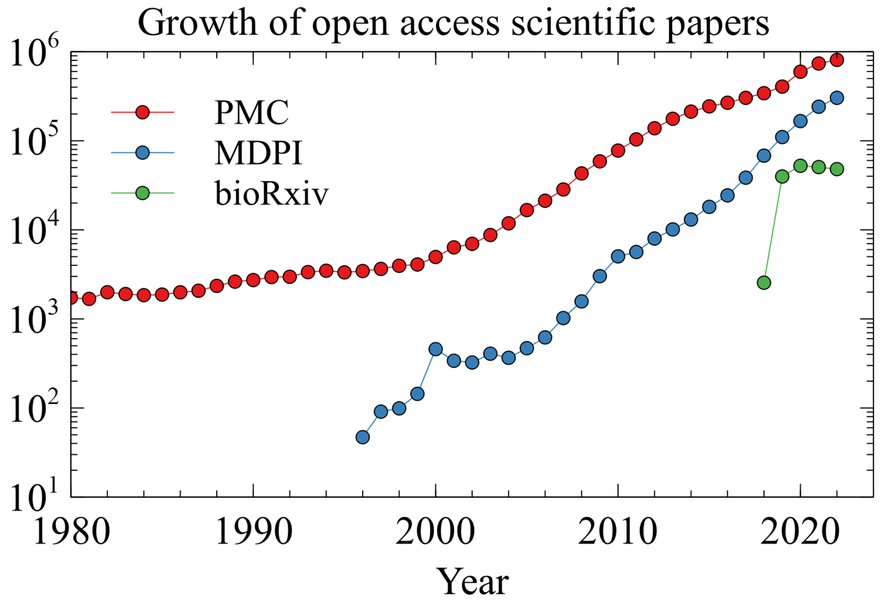

Fig 1. Paper growth in repositories of open access papers: (1) PubMed Central, (2) Multidisciplinary Digital Publishing Institute and (3) The preprint server for biology - bioRxiv

Solution: Karel 3.0

Scientific Paper Analyzer (aka Karel 3.0) pipeline:

Question

1. Corpus of scientific papers

- DOI search + document fetch
- Direct PDF upload
- Google Scholar query search + document fetch

2. Processing

- GROBID
  - PDF to XML conversion
  - Contextual paragraph chunking
- Filter (discard bad/malformed chunks)
- Chunks embedding using
  - multi-qa-mpnet-base-dot-v1
  - multi-qa-mpnet-base-cos-v1

3. Semantic search

- Faiss (Facebook AI similarity search) for relevanat chunks
- Different pipeline is used when full text operation is detected
- Optionally also search all previously uploaded papers
  - plus pool (10M) of already embedded high-quality papers

4. Reranking

- With respect to the question using 'ms-marco-MiniLM-L-6-v2'
- Sentence transformers
- Context expansion (include spatially adjacent chunks)

5. Question answering

- GPT 4.0 (preprompted to use only provided context)
- Answers are generated per paper to keep reference

Answers

Results and Future Work

Currently, our service is a fully operational unit and supports all publicly available scientific papers receiving highly positive feedback from our researchers. They have noted the tool's efficiency in reducing the time and effort needed to find relevant literature, making it an essential asset for accelerating research.

Looking ahead, we plan to further enhance our system by integrating a local LLM and replacing the current GPT usage with 'Mistral-7B-Instruct', a smaller yet highly effective model (we conducted an extensive model evaluation for this particular use case). This shift will improve processing speeds and data privacy.

The development of Karel took approximately 2 man-years, with an additional 1 man-year equivalent in on-premise hardware costs.

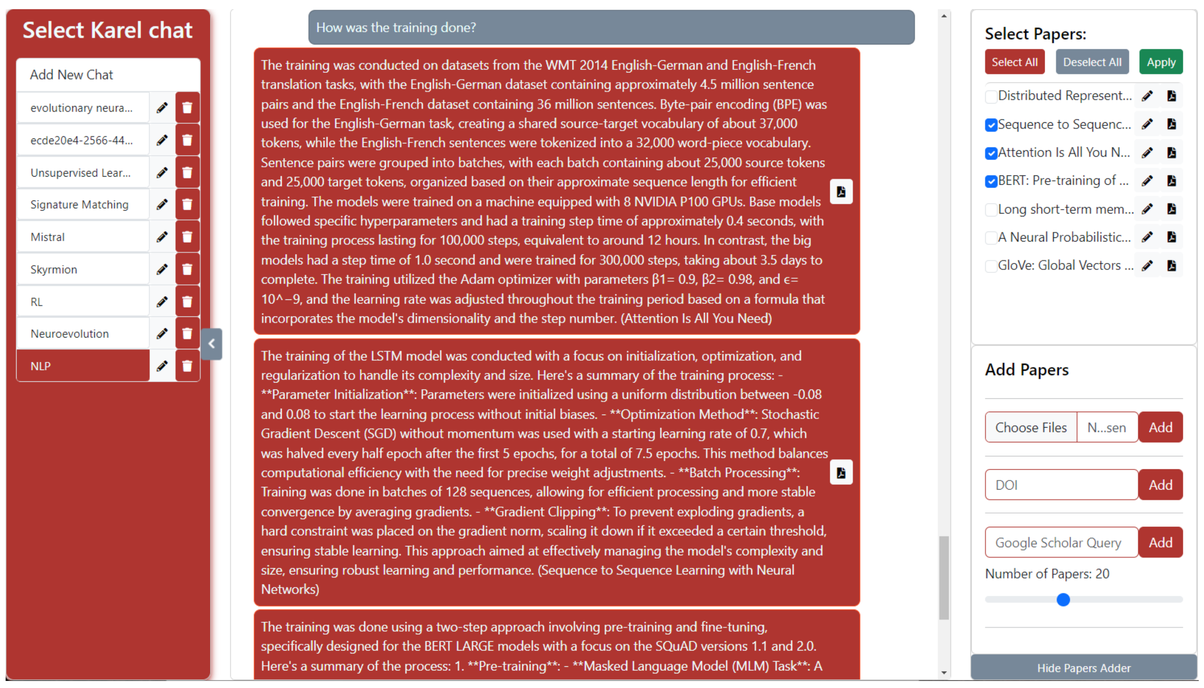

Fig 2. Karel 3.0 user interface

Bene Meat Technologies a. s. is a Czech startup focused on development of cultivated meat. On 8. 11. 2023 it became the first in the world to obtain registration of a cultivated meat product for pet feed.