NLP-BASED COGNITIVE SEARCH ENGINE FOR THE GEOSS PLATFORM DATA

Exploiting existing GEOSS datasets for Climate Change adaption and mitigation

Yannis Kopsinis*, Zisis Flokas*,
Pantelis Mitropoulos*, Christos Petrou*, Thodoris Siozos*, Giorgos Siokas*
**LIBRA AI Technologies, yannis.kopsinis@libramli.ai*

Scan for abstract

Sharing is encouraged

LIBRA A.I. Technologies

Eiffel GEOSS APPLICATIONS FOR CLIMATE CHANGE

01. Introduction

This work presents a domain-aware cognitive search engine (SE) designed in EIFFEL H2020 project. It aims to exploit recent advances in Machine Learning (ML)-based Natural Language Processing (NLP) to overcome current challenges in the searching capabilities of Data Portals. The system includes an optimized AI Large Language Model (LLM) retrained with an extensive Climate Change (CC)-specific text corpus. Cognitive search adds language understanding to the search results, promoting the most semantically relevant results to the top. The use-case is the GEOSS Portal, but the same principles apply elsewhere.

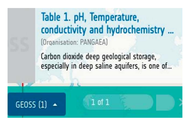
02. Conventional vs Cognitive search data discovery experience

Conventional search engines, facts

- Based on exact or fussy term searches.
- Limited data discovery capabilities.
- The inherent limitations of such approaches are expressed in a higher degree in metadata querying since the available text is limited to, usually, title, description and keyword lists.

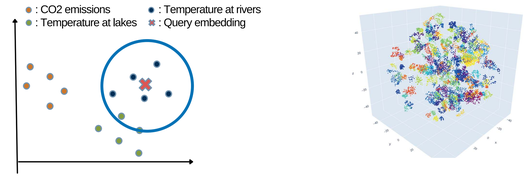
Cognitive search engines, facts

- Based on data-driven language models.
- Allow free-text querying.
- The language model inherently performs the semantic analysis: It understands individual words' meanings, the meaning of words within their context, the semantic relationships between individual words and even the meaning of whole sentences.

Original queries	*Alternative equivalent queries*	*Cognitive Search Results*	
Agricultural pollution to rivers	*Pollution to rivers due to agriculture*	*Agricultural pollution to rivers* (and all 4 alternatives)	*Greenhouse gases emissions* (and alternative spelling)
	 The SE needs to be resilient to rephrased queries	**Annual CO2 emissions from regulated installations** Annual emissions of carbon dioxide equivalent from installations in England holding a Greenhouse Gas Emissions Permit under the... **Annual CO2 emissions from aircraft operators** Annual emissions of carbon dioxide from those aircraft operators regulated under EU Emissions Trading Scheme and assigned to the UK... **Methane, liquid petroleum gas, smoke, carbon monoxide and...** Methane, liquid petroleum gas, smoke, carbon monoxide and propan obtained during the pandemic period in Ankara, Turkey **Green house gas and meteorology data obtained during pandemic...** In this study a greenhouse gas and meteorology measurement station is developed to monitor ozone, methane, ammonia, nitrogen dioxide.... **Copernicus Atmosphere Service near real-time biomass burning...** This service provides pre-operational daily analyses of biomass burning emissions based on fire radiative power satellite...	**Diffuse agricultural pollution to rivers in Scotland** Modelled predictions of annual pollutant loads in rivers from agricultural source areas for... **Water footprint of irrigated agriculture in Segura river basin...** Irrigated agriculture is a key activity in water resources management at the river basin level in arid and semi-arid areas, since this sector... **Environmental Pollution Incidents** Environmental Pollution Incident data filtered for Categories 1 and 2. Details of all pollution incidents reported to the Environment Agency... **Discharges to rivers from abandoned metal mines** Information about average flows and water quality for known mine water discharges from abandoned metal mines in England... **Baseflow chemistry of streams draining rural and agricultural...** Data from a water quality survey of streams draining rural and agricultural land to the North Sea from the Wolds region of the West...
Pollution AND rivers AND agriculture	*Pollution AND inland waters AND agriculture*		
	 The SE needs to tackle Semantically similar words and concepts consistently.		
Greenhouse gases emissions	*Greenhouse gasses emissions*		
	 The search engine needs to be resilient to misspellings.		

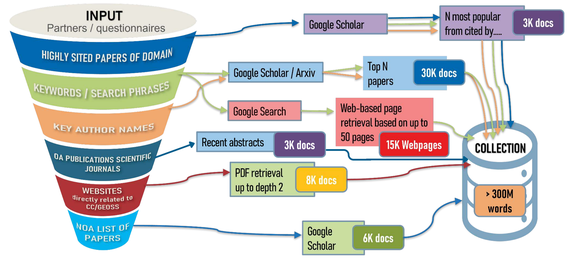
03. How does cognitive search works

- The LLM transforms all documents into mathematical vectors (**Embeddings**), inherently performing semantic analysis.
- Words/terms and sentences with similar concepts and meanings lie close in the embedding space.
- The distance between the document and the query embeddings measures the relevance between a document in the database and any query.
- Semantic search adds language understanding to search results, promoting the most semantically relevant results to the top.
- It can be domain-aware: In EIFFEL we aim for CC domain specificity.



Text embedding encodes words and sentences as numeric vectors

04. CC Domain-specific corpus collection for LLM training

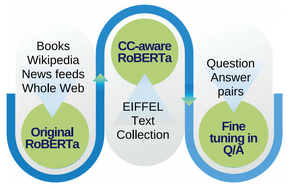


"basin", "distribution", "parameters", "factors", "regions", "environmental", "variables", "emissions", "simulation", "atmospheric", "correlation", "modelling", "measurement", "estimation", "greenhouse", "radiation", "percentage", "climatic", "cooling", "rainfall", "regression", "gases", "pollution", "meteorological", "dioxide", "flux", "anthropogenic", "indicator", "humidity", "ocean", "baseline", "ecosystems", "renewable", "hydrological", "sustainable", "socioeconomic", "CO_2"

CC-related document collection process for LLM retraining (13M sentence) and newly included terms in the Large Language Model

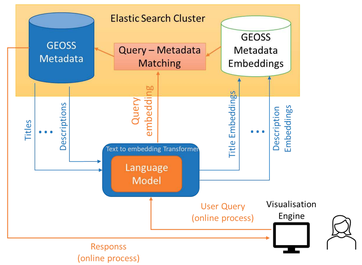
05. LLM fine-tuning for domain-aware cognitive search

- **Step A**: Unsupervised learning with the CC-related corpus that includes new terms[1].
- **Step B:**
  - We use an independent, instruction-based LLM (such as chatGPT) to generate Q&A pairs from the CC-related corpus and the GEOSS Portal metadata description field.
  - We use this new dataset for finetuning in the domain using Generative Pseudo Labeling (GPL)[2] approach.
- **Alternative path:** The Q&A pairs dataset is used for supervised training of a dedicated Cross-encoder (work in progress).



06. EIFFEL Cognitive search pipeline

- The metadata (e.g., titles, descriptions, keywords) pass through the LLM to produce metadata embeddings (offline process).
- The user query passes through the language model to produce the query embedding (online process).
- The semantically similar data objects are returned in ranked order.
- Elasticsearch stores embeddings and calculates vector similarity fast.



References

1. *Wang, Kexin, Nils Reimers, and Iryna Gurevych. "TSDAE: Using Transformer-based Sequential Denoising Auto-Encoderfor Unsupervised Sentence Embedding Learning." Findings of the Association for Computational Linguistics: EMNLP 2021.*
2. *Wang, Kexin, et al. "GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval." Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Acknowledgment

This work is part of the EIFFEL EU project. The EIFFEL project has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101003518.