TECHNISCHE
UNIVERSITÄT
DRESDEN

Semantic Search for Climate Data

Design of a vertical search engine based on Neural Networks

Simeon Wetzel, simeon.wetzel@tu-dresden.de, TU Dresden, Chair for Geoinformatics, Helmholtzstraße 10, 01069 Dresden, Germany

Background & Objectives

Traditional information retrieval approaches involve **lexical matching** of queries with keywords present in the **metadata**. These methods have limitations, such as the inability to handle queries with **synonyms**, abbreviations, or misspelled user inputs. Also the accuracy of search highly depends on the **completeness of metadata [1]**. In contrast to data in geospatial catalogues, climate data provided online often lacks proper metadata.

There has been significant progress in using **NLP (Natural Language Processing)** models to generate **semantically meaningful** and **context aware** embeddings **[2]**. Semantic search embeds both queries and indexed documents into a shared dense vector space and ranks relevance of results based on closeness within this vector space. The objective of this work is to fit these encoder models for the usage in a vertical search engine for climate data.

In order to perform well in the target domain the models have to be **domain adapted** with the specific vocabulary **[3]**. The final encoding model shall also be trained to represent **spatial** and **temporal** context of user queries with seperate models (Fig.1, spatial / temporal branch). The outcome is a search engine that is able to rank search results mainly based on its textual description and is therefore less depended on the completeness and accuracy of metadata.

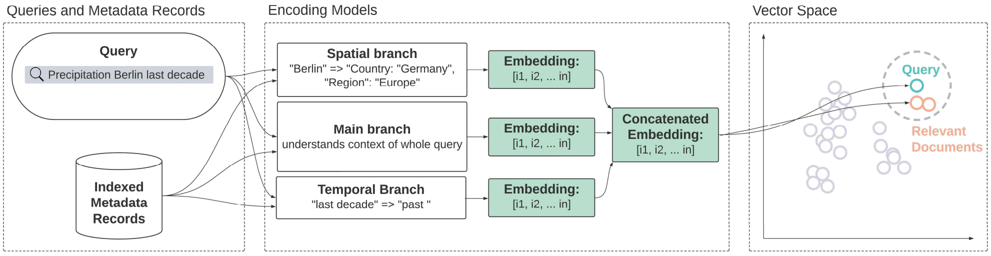
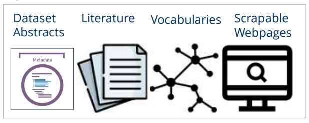


Fig. 1

Methods

Model **training** usually requires a **huge amount** of labelled text that does not exist for the specific domain of climate data. The main branch model is trained with an **unsupervised training** method proposed by **[4]** that only requires an **unlabeled** domain specific **text corpus**. An own corpus was designed with text from multiple sources such as **data abstracts** from **metadata catalogues** (e.g. EEA SDI catalogue), **literature, controlled vocabularies**, and **scrapable webpages** (e.g. Wikipedia) (Fig. 2). The upcoming tasks involve training models for the spatial/temporal branches and an evaluation of the overall performance. The authors of **[5]** proposed a training approach for retrieval models that incorporates **geographical proximity** between spatial entities in both **queries** and retrieved **passages**. The **spatial branch** model shall be trained with a modified version of **[5]**. With our approach, a corpus is offered in which all spatial entities (e.g., city names) are **substituted** with **placeholders** (Berlin => [CITY]) (see Fig. 3).

Fig. 2



Next, spatial entity names obtained from a **gazetteer** are **randomly inserted**, specifically targeting the **geographical region** for which the search is intended ([CITY] => London). That offers the flexibility to **re-utilize** the **corpus** for different regions as needed, thus preventing the underrepresentation of certain geographic regions.

The **temporal branch** a model is a **Named Entity Recognition** (NER) model that classifies queries and passages into ***past*** (e.g. observation data), ***present*** and ***future*** (e.g. climate projections).

To **evaluate** the search, it is planned to obtain **relevance scores** for a curated set of queries and passages through expert interviews. Then, common metrics, such as the normalized Discounted Cumulative Gain (**nDCG**) can be applied to evaluate the model performance.

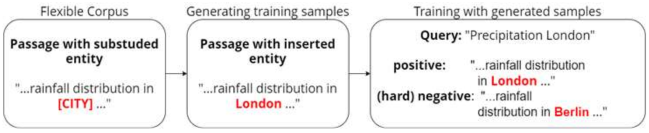

Flexible Corpus
Passage with substuded entity
"...rainfall distribution in [CITY] ..."
Generating training samples
Passage with inserted entity
"...rainfall distribution in London ..."
Training with generated samples
Query: "Precipitation London"
positive: "...rainfall distribution in London ..."
(hard) negative: "...rainfall distribution in Berlin ..."

Fig. 3

Results & Outlook

The presented models and architecture can be **implemented** into **the search backends** of common data repositories such as **CKAN** or **GeoNetwork**. We set up a **prototype** CKAN and harvested ~106k records from the **World Data Centre for Climate** (https://www.wdc-climate.de/ui/) to test the search. Fig 4. shows first tests with example queries and the top 50 ranked records retrieved using the standard BM25 and the trained dense retriever model. The tests show a higher amount of relevant records (true positive) retrieved by the dense retriever than BM25, but also with a lower precision (also more false positive records).

Fig. 4

Q1: „drought return rate“

	BM25	AI-model
True Positive	1	42
False Positive	0	8

Q2: „adaptation measures “

	BM25	AI-model
True Positive	1	4
False Positive	1	13

Q3: „precipitation projection“

	BM25	AI-model
True Positive	50	50
False Positive	0	0

Q4: „cool days“

	BM25	AI-model
True Positive	0	20
False Positive	0	30

Figures:

Fig. 1: Schematic view of the search engine architecture
Fig. 2: Corpus sources for domain adaption
Fig. 3: Workflow for generating customized training samples for the spatial search
Fig. 4: Example queries and recieved relevant/irrelevant results under the top 50 ranked documents

Literature:

1 *Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L. D., Kacprzak, E., & Groth, P. (2020). Dataset search: a survey. VLDB Journal, 29(1), 251–272. https://doi.org/10.1007/s00778-019-00564-x*

2 *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-Decem(Nips), 5999–6009.*

3 *Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeurIPS. http://arxiv.org/abs/2104.08663*

4 *Wang, K., Thakur, N., Reimers, N., & Gurevych, I. (2021). GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. http://arxiv.org/abs/2112.07577*

5 *Coelho, J., Magalhães, J., & Martins, B. (2021). Improving Neural Models for the Retrieval of Relevant Passages to Geographical Queries. GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, 268–277. https://doi.org/10.1145/3474717.3483960*