EGU24-1707, updated on 22 Jul 2024
https://doi.org/10.5194/egusphere-egu24-1707
EGU General Assembly 2024

Topic Analysis and Classification of EGU Conference Abstracts

Jens Klump[1,3], Chau Nguyen[2,3], John Hille[1,3], and Michael Stewart[1,2,3]
[1]CSIRO, Mineral Resources, Kensington, Australia (jens.klump@csiro.au)
[2]University of Western Australia, Perth, Australia
[3]Centre for Transforming Maintenance through Data Science, Perth, Australia

The corpus of Abstracts from the EGU General Assemblies 2000 - 2023 covers a wide range of Earth, planetary and space sciences topics, each with multiple subtopics. The abstracts are all in English, fairly uniform in length, cover one broad subject area, and are licenced under a permissive licence that allows further processing (CC BY 4.0), making this a high-quality text corpus for studies using natural language processing (NLP) and for the finetuning of Large Language Models (LLM). Our study makes use of openly available NLP software libraries and LLMs.

In the first phase of this study, we were interested in finding out how well abstracts map to the topics covered by EGU Divisions and whether co-organisation of sessions contributes to or dilutes topics. The abstracts are available only in unstructured formats such as Portable Document Format (PDF) or plain text in XML extracts from the conference database. They are identified by abstract numbers but carry no information on the session or division where they were originally presented. We reconstructed this information from the online conference programme.

To be able to employ a supervised learning approach of matching abstracts to topics, we defined the topics to be synonymous with the 23 scientific divisions of the EGU, using the division and co-listed divisions as topic labels.

We finetuned the Bidirectional Encoder Representations from Transformers (BERT) and the slightly simplified DistillBERT language models for our topic modelling exercise. We also compared the machine classifications against a random association of abstracts and topics. Preliminary results obtained from our experiments show that using a machine learning model performs well in classifying the conference abstracts (accuracy = 0.66). The accuracy varies between divisions (0.40 for NP to 0.96 for G) and improves when taking co-organisation between divisions into account. Starting from one year of abstracts (EGU 2015), we plan to expand our analysis to cover all abstracts from all EGU General Assemblies (EGU 2000 - 2024).