

# Advanced IR Experiments with Language Models, Summarization and Translation using SBERT, Elasticsearch and trec\_eval

## Group 05

David Mihola ([david.mihola@student.tugraz.at](mailto:david.mihola@student.tugraz.at), 12211951), role: Task 01

Manuel Riedl ([manuel.riedl@student.tugraz.at](mailto:manuel.riedl@student.tugraz.at), 12114848), role: Task 02

Massimiliano Viola ([massimiliano.viola@student.tugraz.at](mailto:massimiliano.viola@student.tugraz.at), 12213195), role: Task 03

Nico Ohler, ([nico.ohler@student.tugraz.at](mailto:nico.ohler@student.tugraz.at), 12118907), role: Task 04

<https://github.com/massimilianoviola/advanced-information-retrieval>

# Motivation

- How do neural IR and sentence transformers perform compared to traditional approaches?
- Can we save storage space by using summarized documents, while still maintaining or even improving search performance?
- Can we make documents accessible to different languages without translation negatively impacting search?
- What are the search implications of using a multilingual model when handling queries in multiple languages? Are they viable alternatives to their monolingual counterparts?

# Datasets

Three datasets of different domains and different sizes were employed to test pattern generalization in the results:

Medline:

- domain-specific records from a medical journal
- 1033 long and analytical articles
- 30 queries

CACM:

- 3204 publication records
- significant fraction also having an abstract
- 64 queries

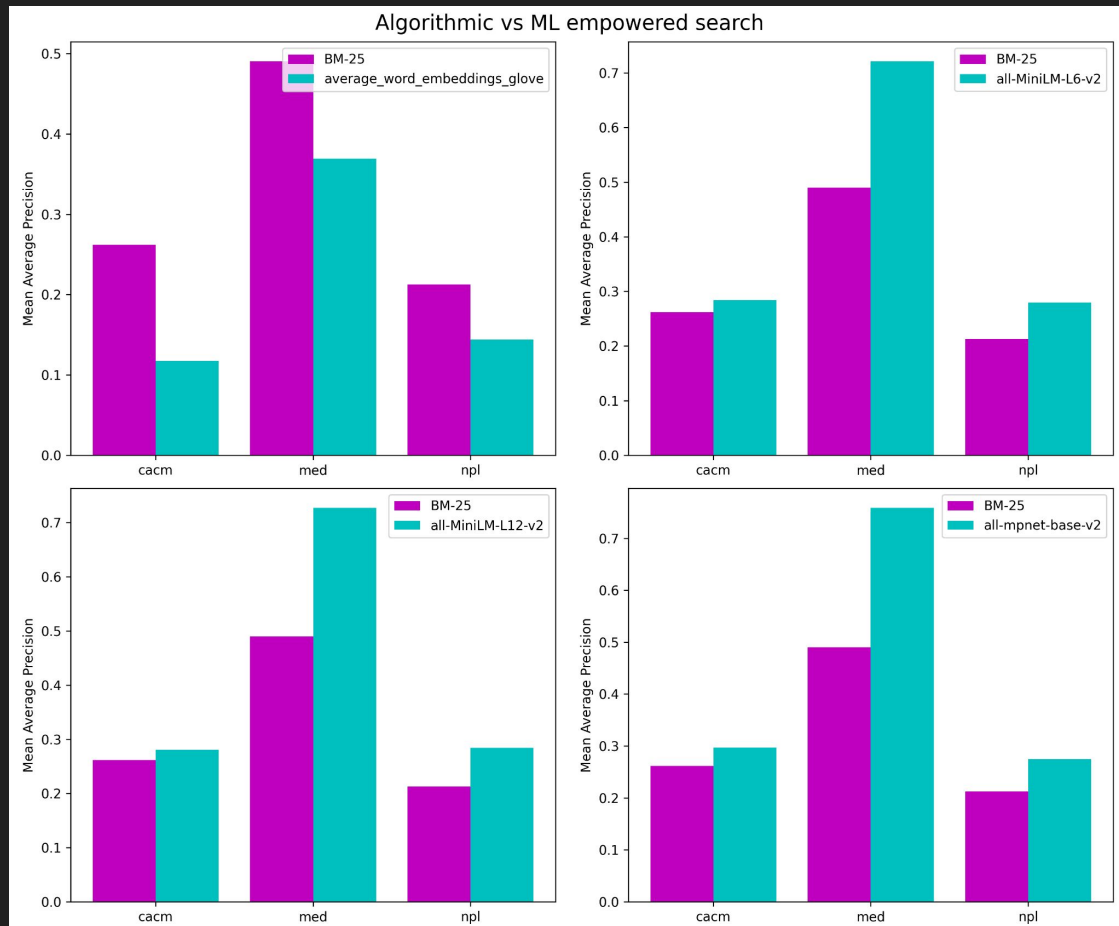
NPL:

- 11429 relatively short document titles
- 93 queries

# Tools

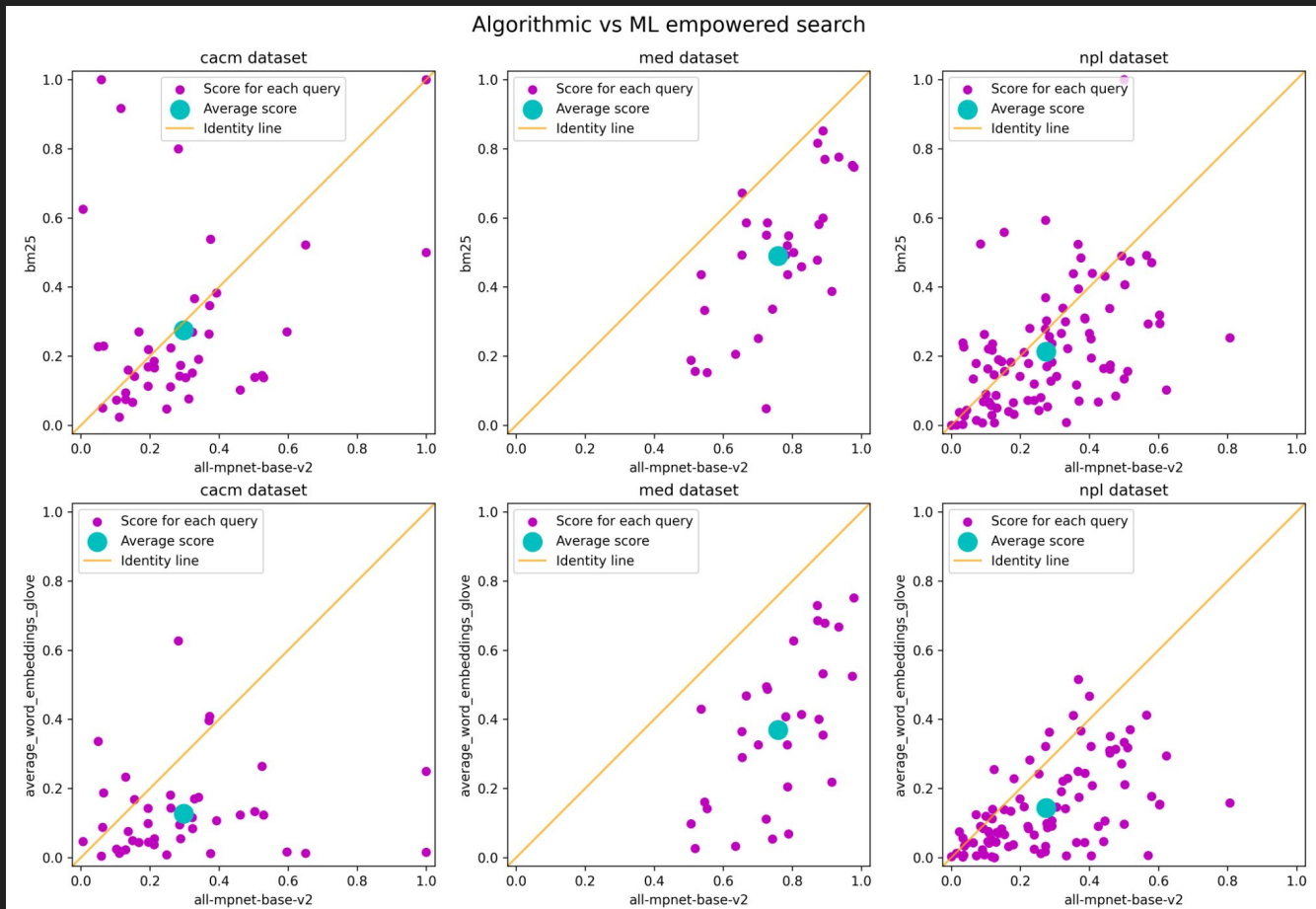
- Elasticsearch
  - full-text search engine
- trec\_eval
  - evaluate text retrieval results
- BM25
  - benchmark relevance ranking function
- SBERT - Sentence Transformers
  - create document embeddings
  - improve contextual understanding
  - rely on attention mechanism
- Hugging Face Transformers
  - text summarization

# Task 01: Sentence Transformers



# Task 01: Sentence Transformers

Algorithmic vs ML empowered search



## Task 02: Summarized vs Non-Summarized

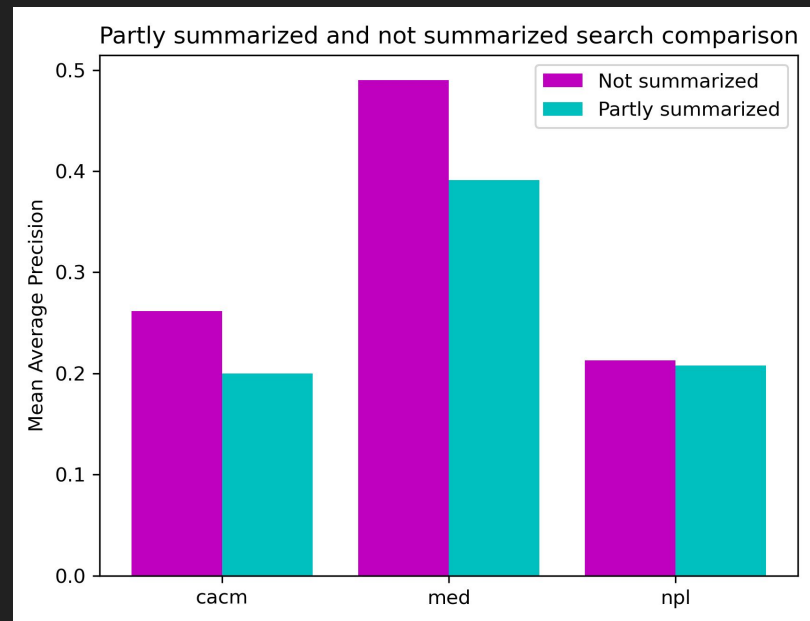
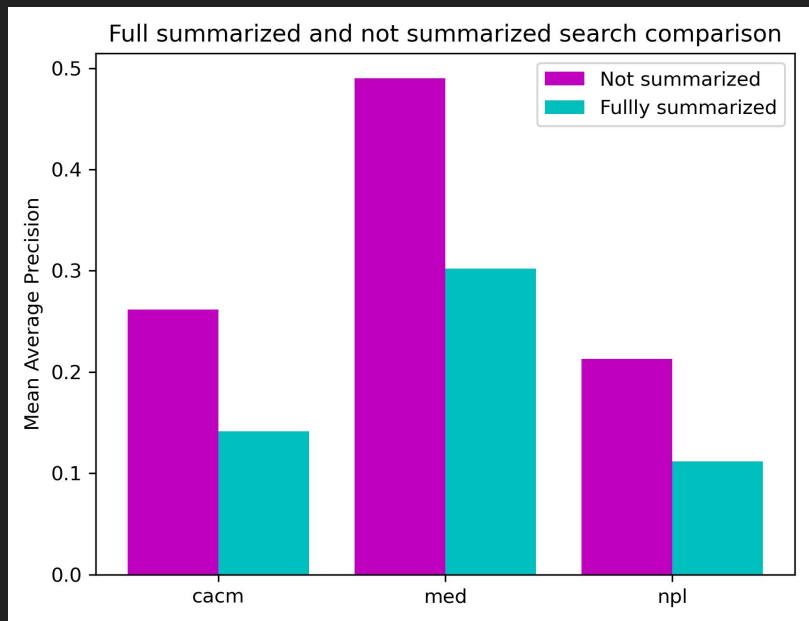
Challenge: text summarization models may lead to strange results if the sequence is too short

"Secant Modification of Newton's Method"



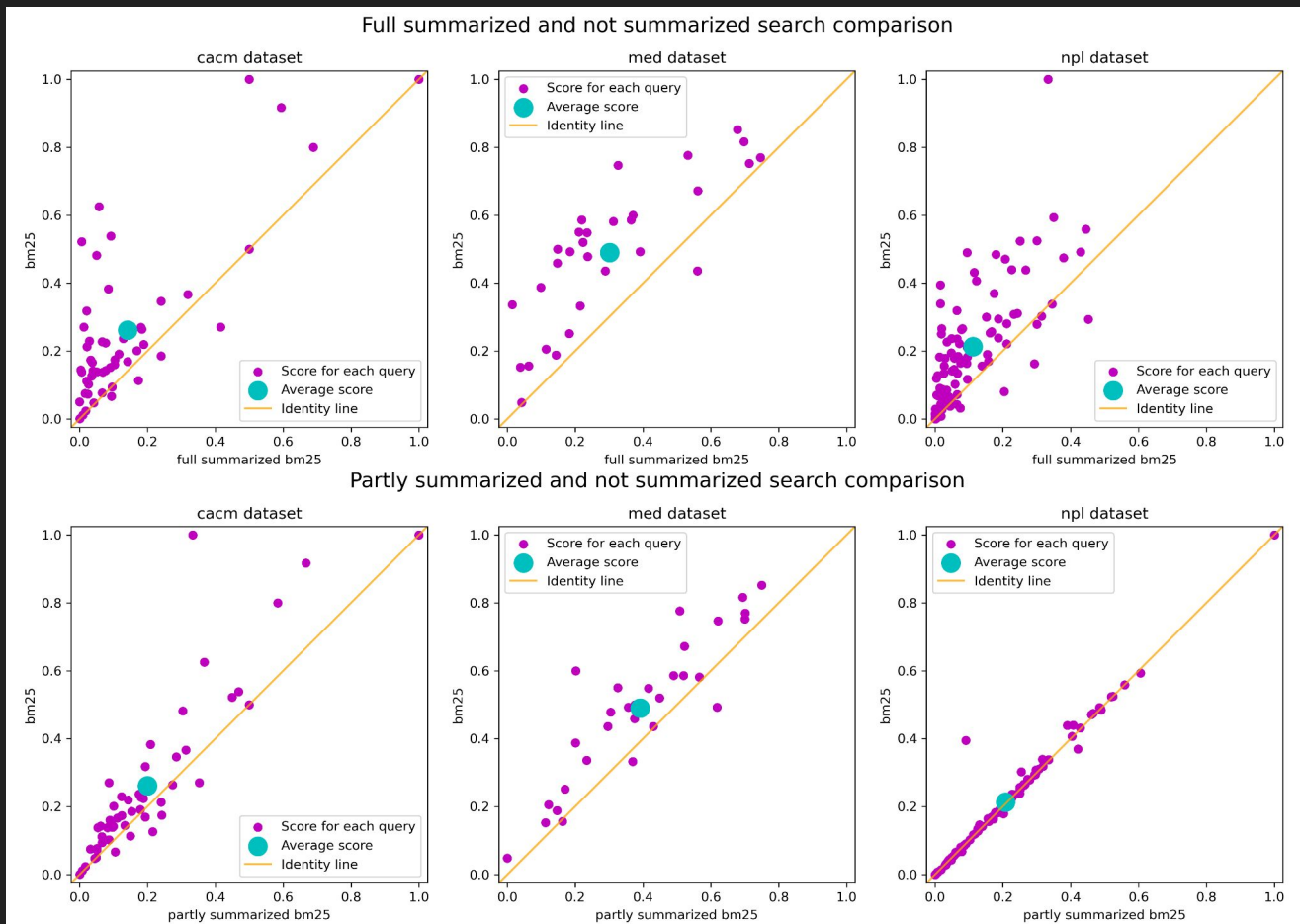
"BBC Sport takes a look at some of the more unusual ways  
in which Newton's method has been used."

# Task 02: Summarized vs Non-Summarized

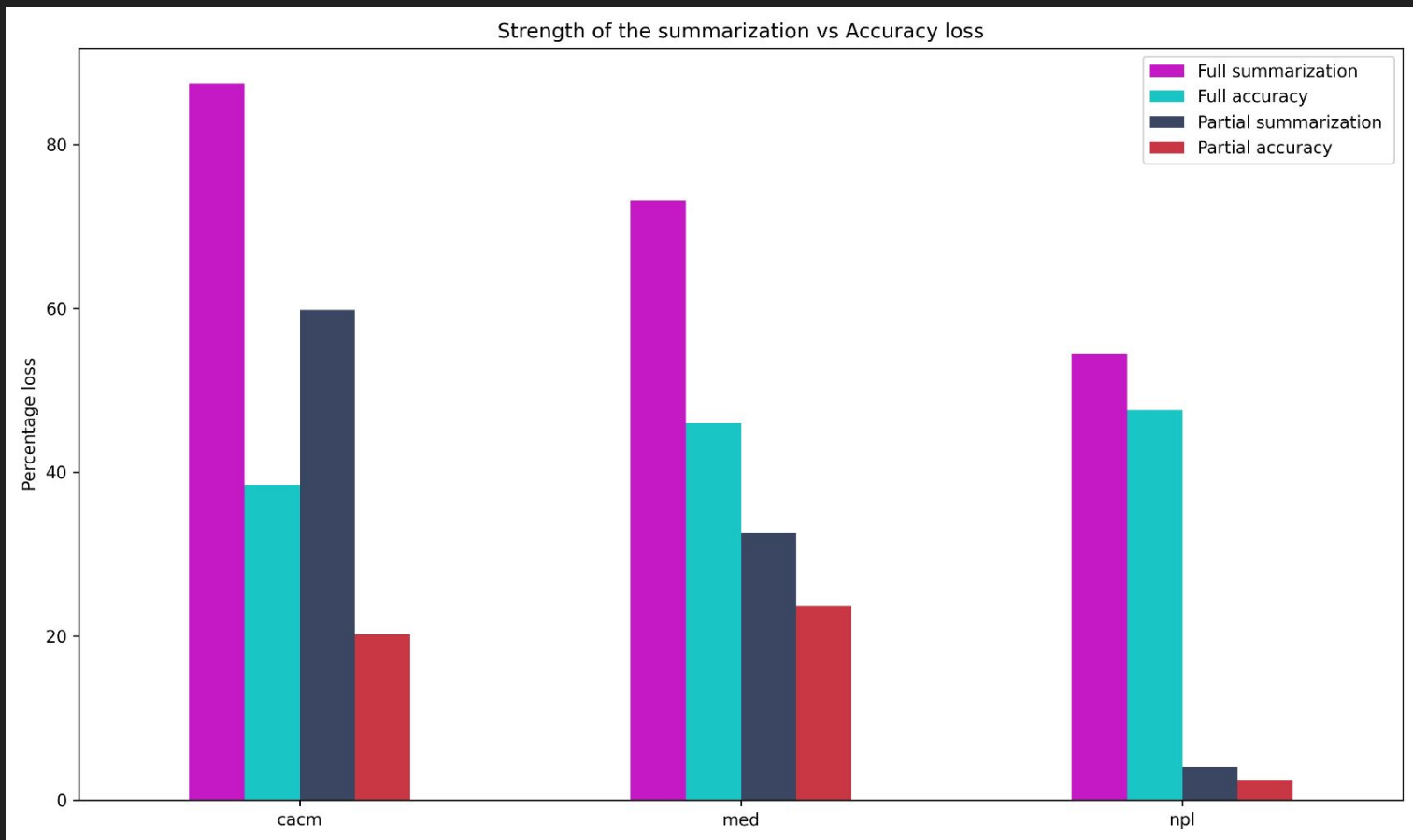




# Task 02: Summarized vs Non-Summarized

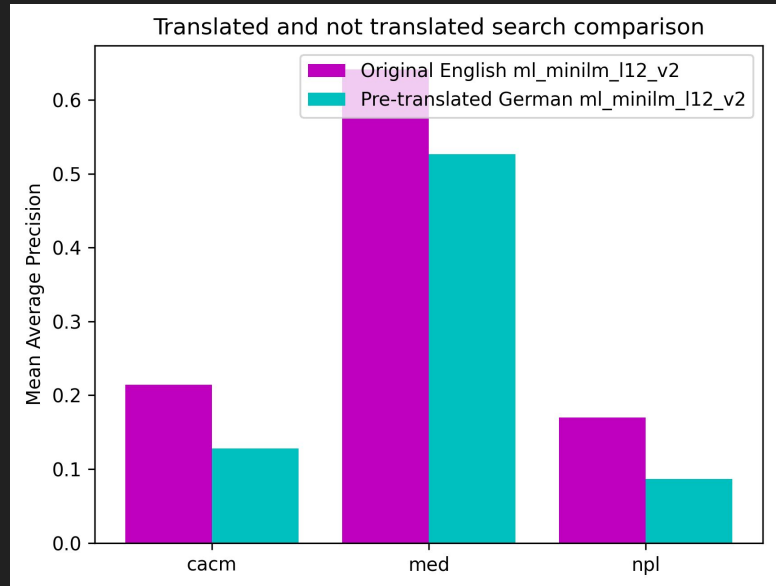
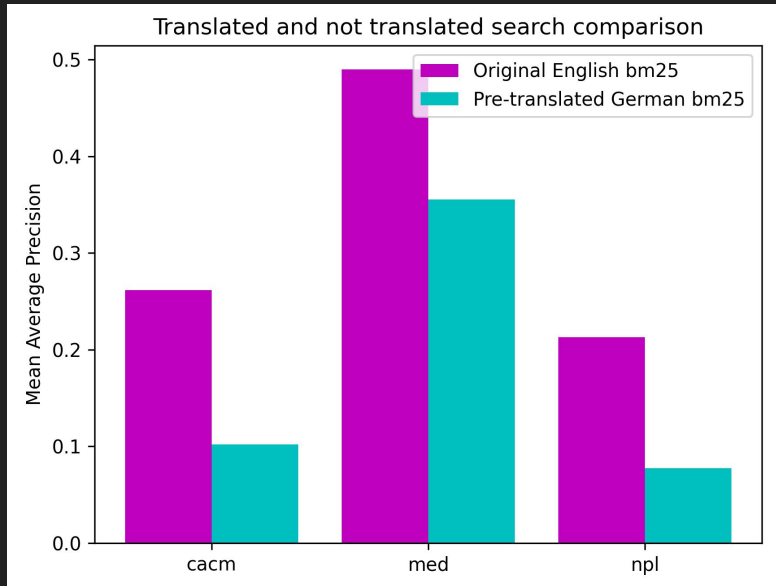


# Task 02: Summarized vs Non-Summarized

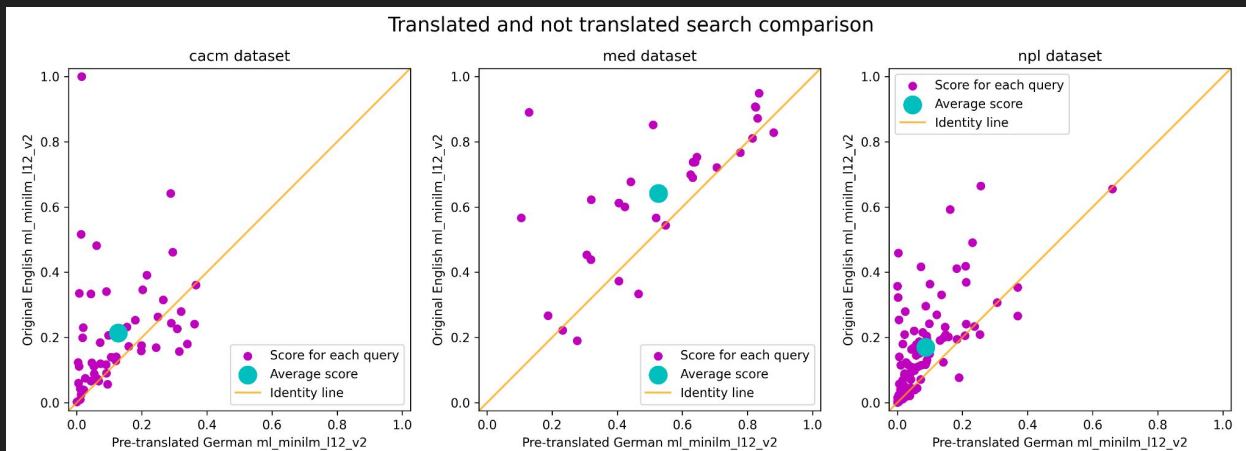
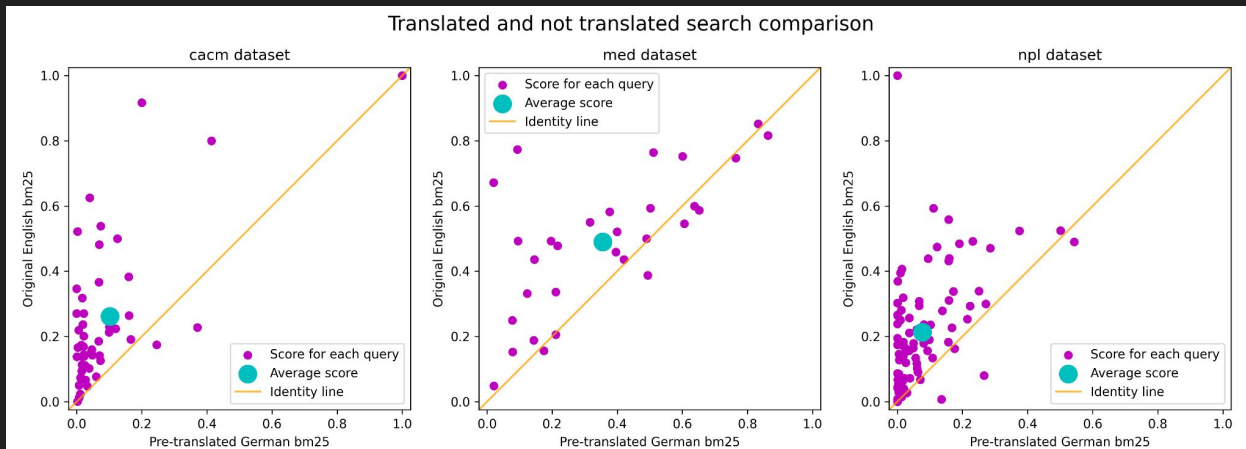


# Task 03: Translated vs Untranslated

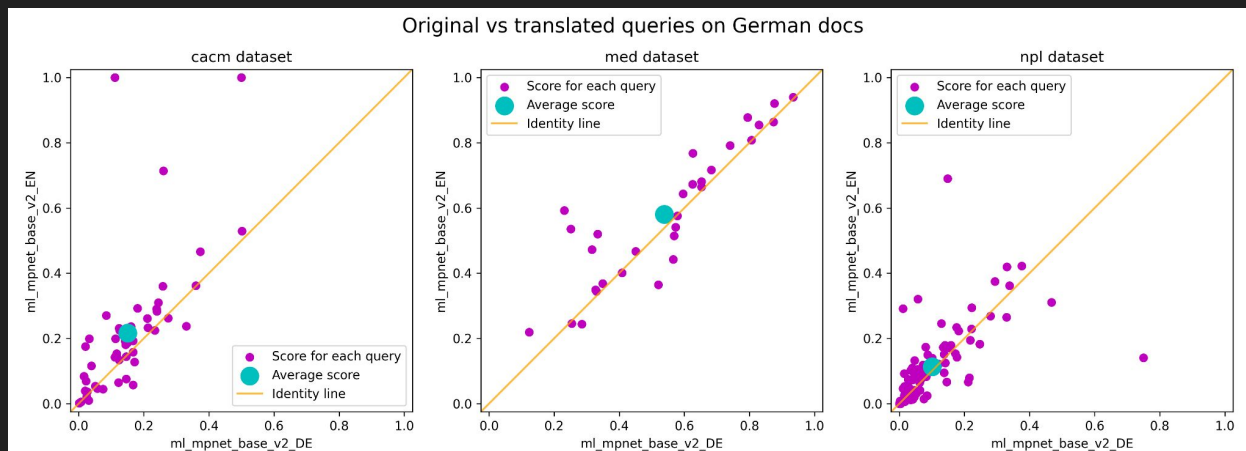
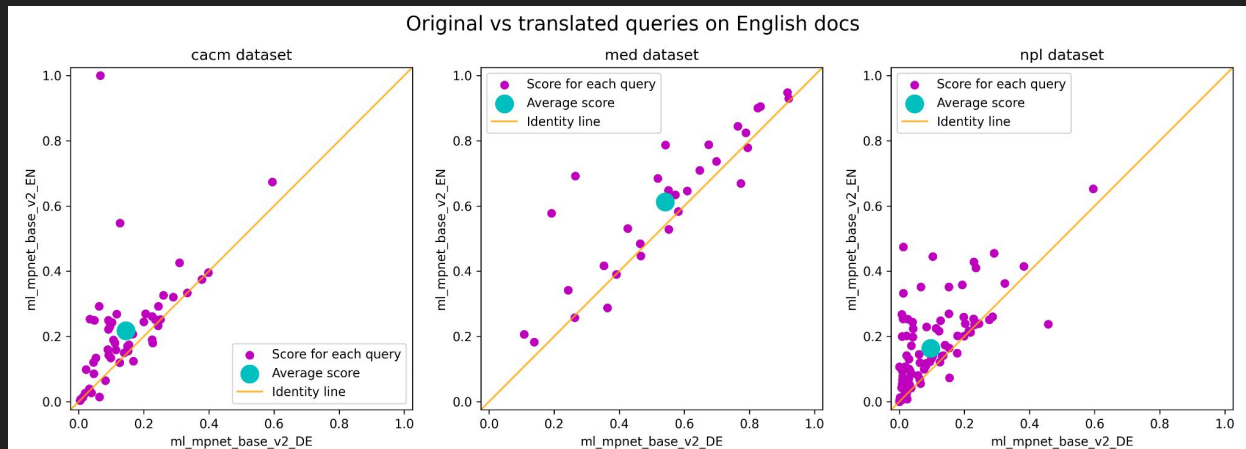
Documents in a minority language may need to be translated into another, more popular language (like English) to be more accessible. We simulate the process in reverse by translating documents and queries into German.



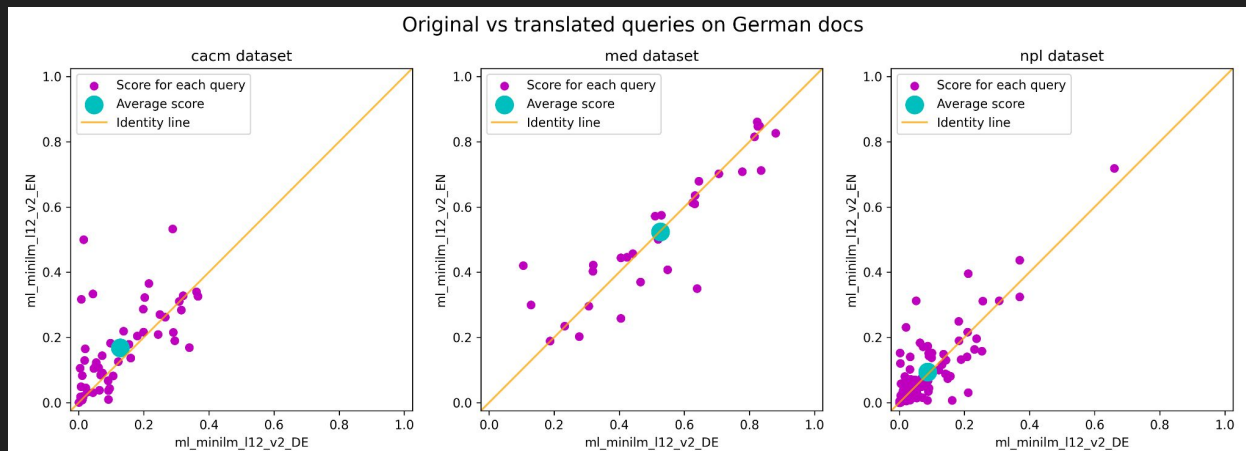
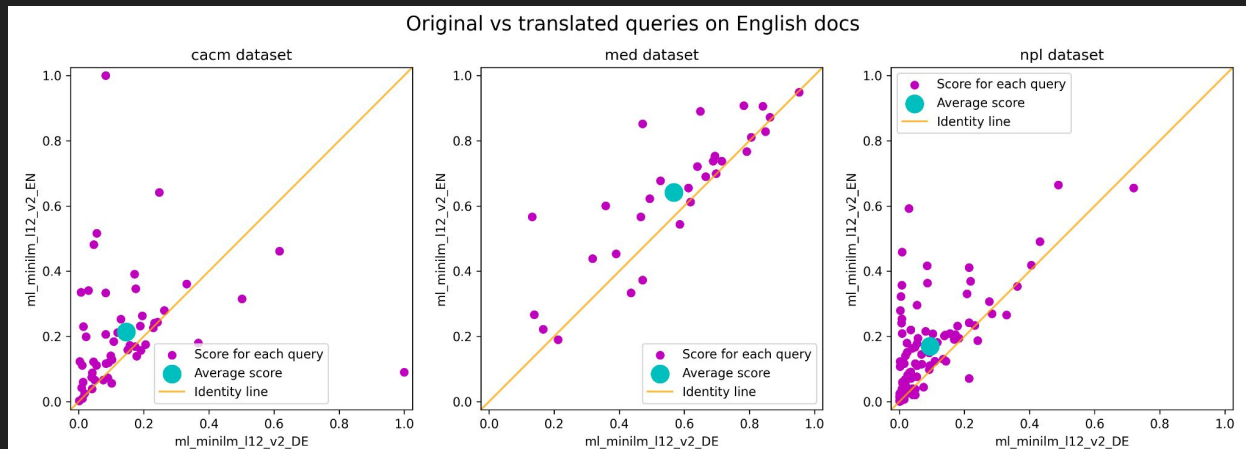
# Task 03: Translated vs Untranslated



# Task 04: Multilingual Sentence Transformers



# Task 04: Multilingual Sentence Transformers



# Conclusions

- Neural IR systems are powerful but very expensive at scale
- BM25 is a strong and fast baseline and can easily beat GloVe
- Text summarization slightly reduces performance and requires caution regarding which documents to summarize depending on text length
- Translation can bring many problems with it
- Multilingual search trades performance for the ability to handle documents and queries in multiple languages