Comparing Different Approaches to cross-lingual Information Retrieval

Group 02

Katharina Aschauer

Distiluse-base Approach
Evaluation
Plotting

Maximilian Binder

mBERT Approach
Experiment Framework
Preprocessing

Jan-Peter Svetits

MiniLM Approach
Presentation
Preprocessing

Github Repository

github.com/katasc22/AIR2023

Introduction

Motivation

Overcome language barriers
Enhance user accessibility

Research Questions

- What is the most convenient approach to retrieve documents from multilingual queries?
- How do different cross-lingual information retrieval approaches perform in comparison to each other?
- What are the tradeoffs between the chosen methods?
- How do multilingual models perform in monolingual settings?

Dataset

Vaswani

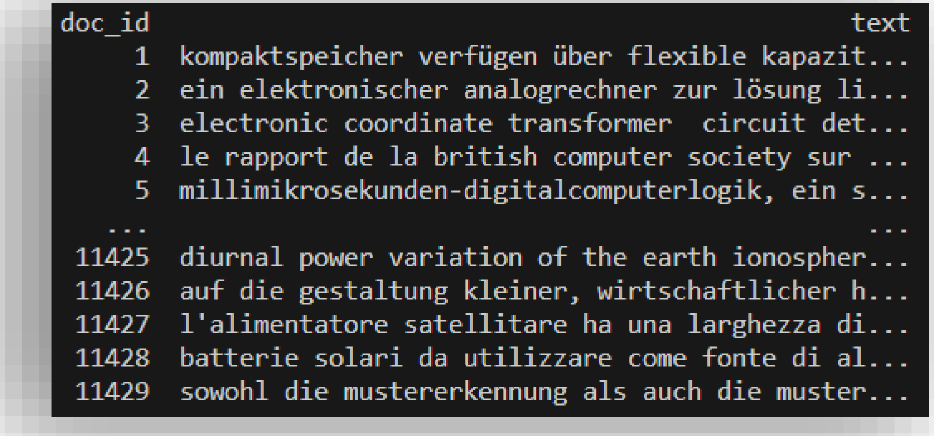
- A small corpus of roughly 11,000 scientific abstract
- 11429 Documents
- 93 Queries
- 2083 Qrels
- English only

Preprocessing

Tasks

- Translation of data into multiple languages

🇬🇧 🇩🇪 🇮🇹 🇫🇷



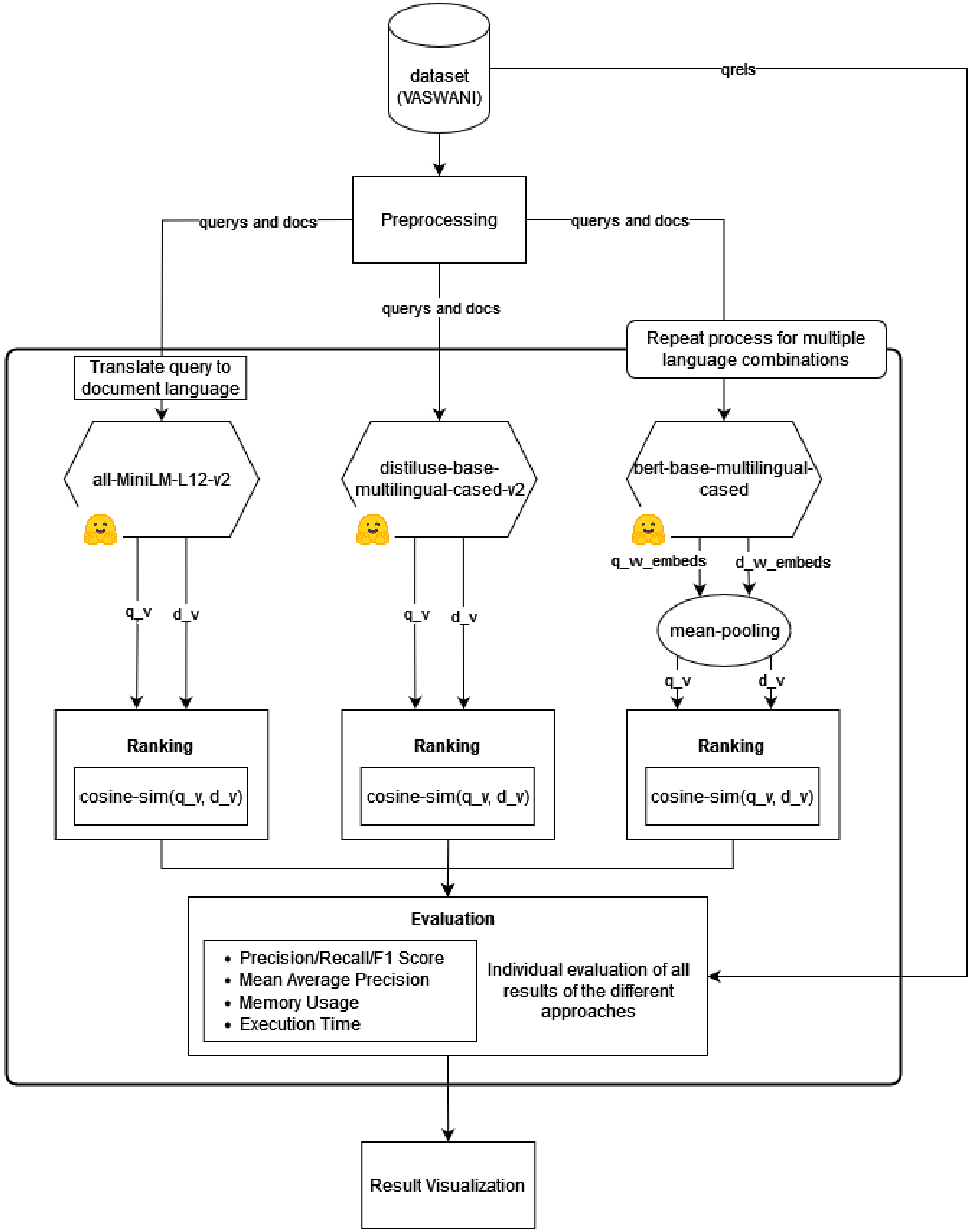
```
doc_id                                                text
     1  kompaktspeicher verfügen über flexible kapazit...
     2  ein elektronischer analogrechner zur lösung li...
     3  electronic coordinate transformer  circuit det...
     4  le rapport de la british computer society sur ...
     5  millimikrosekunden-digitalcomputerlogik, ein s...
   ...                                                 ...
 11425  diurnal power variation of the earth ionospher...
 11426  auf die gestaltung kleiner, wirtschaftlicher h...
 11427  l'alimentatore satellitare ha una larghezza di...
 11428  batterie solari da utilizzare come fonte di al...
 11429  sowohl die mustererkennung als auch die muster...
```

Methods & Models

- **bert-base-multilingual-cased** 🤗
  Pretrained model on the top 104 languages
- **distiluse-base-multilingual-cased-v2** 🤗
  Multilingual knowledge distilled version of multilingual Universal Sentence Encoder with 50+ languages.
- **all-MiniLM-L12-v2** 🤗
  Monolingual model
  (we have to pretranslate the queries before creating embeddings)

Additional Models

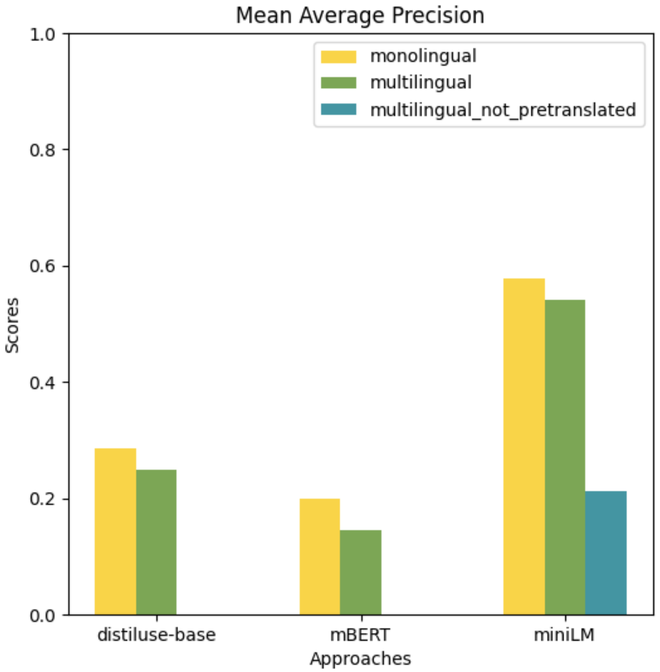
xlm-roberta-base-language-detection 🤗
(for language classification)
Helsinki-NLP/opus-mt-src_lang-target-lang 🤗
(for translation)

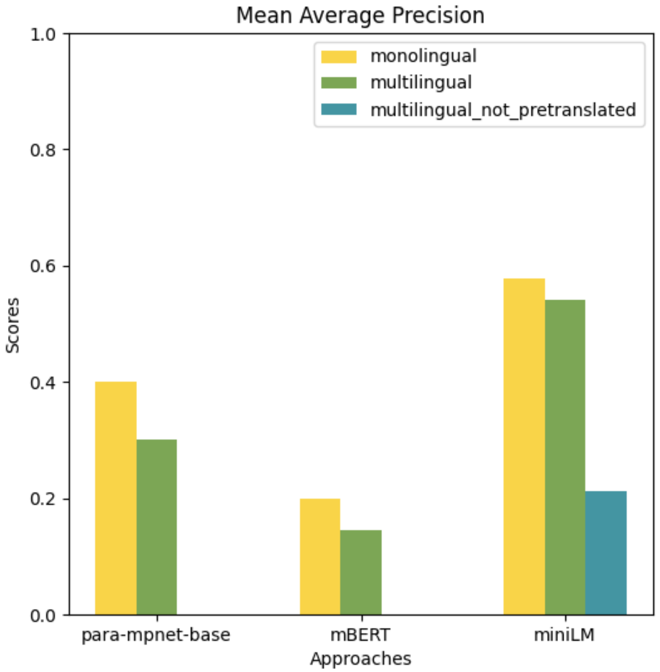


Analysis

Retrieval Performance

- MiniLM has the best scores
  (Monolingual approach)
- Distiluse-base is weaker than expected
- Para-mpnet-base is better for semantic search
  (clear when looking at out-of-the-box semantic search benchmarks for pretrained models in SBERT documentation)
- mBERT has the worst performance





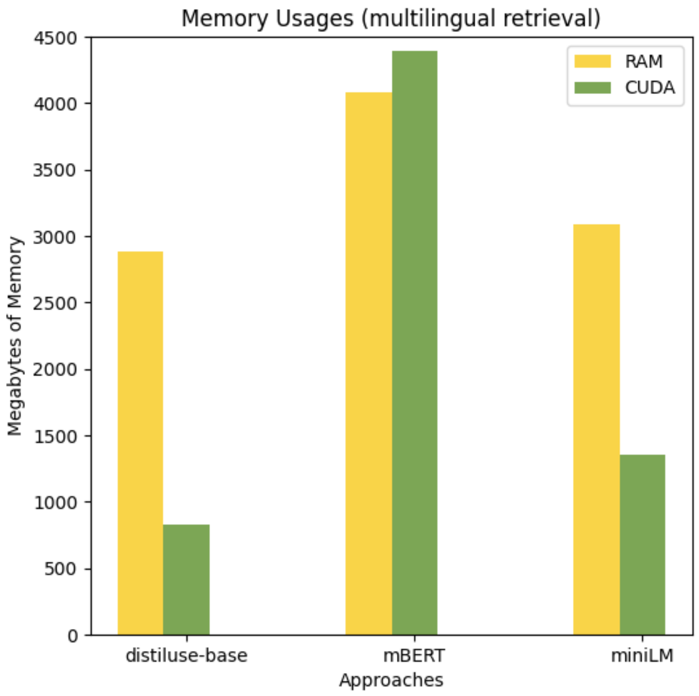
Analysis

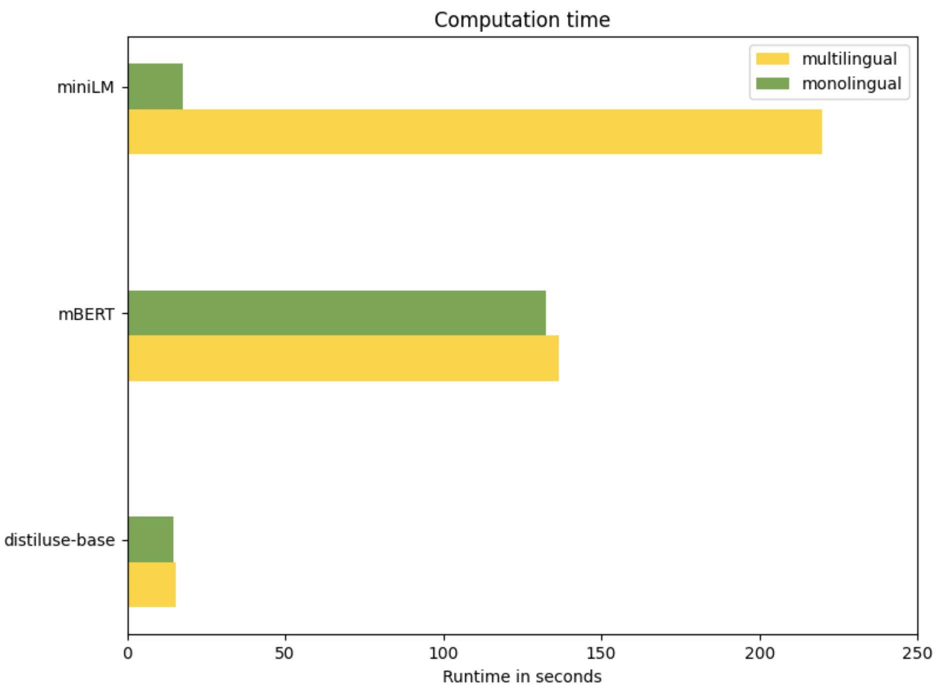
Memory usage

- Distiluse-base and MiniLM use roughly the same memory
- mBERT uses the most memory

Computation time

- MiniLM very slow in multilingual setting
- mBERT bad performance
- Distiluse is the fastest


Memory Usages (multilingual retrieval)


Computation time

Conclusion

Learnings

- Monolingual model with pretranslated queries has good performance but is bad at scale.
- Choosing the right model for the respective task significantly impacts retrieval performance
- mBERT does not produce good sentence embeddings out of the box.

Limitation

- Out-of-box-performance might differ from model to model
- Translation method (for pretranslating text)

Bias

- Dataset might influence results

Best approach

Using fine-tuned multilingual model. It provides the best trade-offs between retrieval performance and computational requirements.