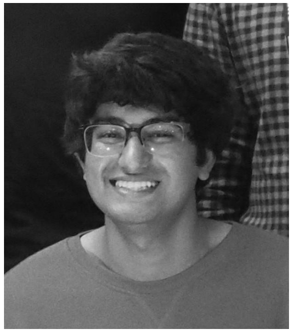
BEIR
An Open-Source Benchmark for Information Retrieval Systems

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych

TECHNISCHE UNIVERSITÄT DARMSTADT





Nandan Thakur
UKP

Nils Reimers
Hugging Face

Andreas Rücklé
Amazon

Abhishek Srivastava, UKP

Iryna Gurevych
UKP

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

https://www.ukp.tu-darmstadt.de/

Beir
Benchmarking IR

UBIQUITOUS KNOWLEDGE PROCESSING

My Journey (Roadmap)

Hi, I'm Nandan!

- I work at an intersection of **NLP**, **Deep Learning** and **Information Retrieval**.
- Currently at UKP Lab, where I am advised by **Dr. Nils Reimers** and **Prof. Iryna Gurevych**.

NLP Researcher
UKP Lab, TU Darmstadt

- I like to study and research on topics with **efficient** and **practical** neural IR.
- I will be starting my PhD soon in September. My advisor will be **Prof. Jimmy Lin**.

Incoming CS PhD
2021-Present, Canada

What we will be learning today?

1. **What is 🔍 Information Retrieval?**

2. **Break down 🔍 Information Retrieval Architecture!**
   2.1 Retrieval Architecture
   2.2 Retrieve and Re-rank Architecture

3. **Traditional Search Systems (BM25)**

4. **Modern Search Systems**
   4.1 Bi-Encoders (Dense Retrieval)
   4.2 Cross-Encoders (Reranking)

5. **Limitations with Traditional and Modern Search Systems!**

6. **Motivate why we create BEIR? Why you should use BEIR?**

7. **Comparison of Search Systems on BEIR!**
   7.1 Performances
   7.2 Efficiency and Speed

8. **Conclusion with Additional Information.**

What is 🔍 Information Retrieval?

🔍 **Which football club Lionel Messi plays for?**

natural language query

OR

🔍 **Messi football club**

keyword-based query

WIKIPEDIA
The Free Encyclopedia

5.5M Articles

Lionel Messi

Lionel Andrés Messi (born 24 June 1987), also known as Leo Messi, is an Argentine professional footballer who plays as a forward for Ligue 1 club **Paris Saint-Germain** and captains the Argentina national team. Often considered the best player in the world and widely regarded as one of the greatest players of all time, Messi has won a record six Ballon d'Or awards, a record six European Golden Shoes, and in 2020 was named to the Ballon d'Or Dream Team.

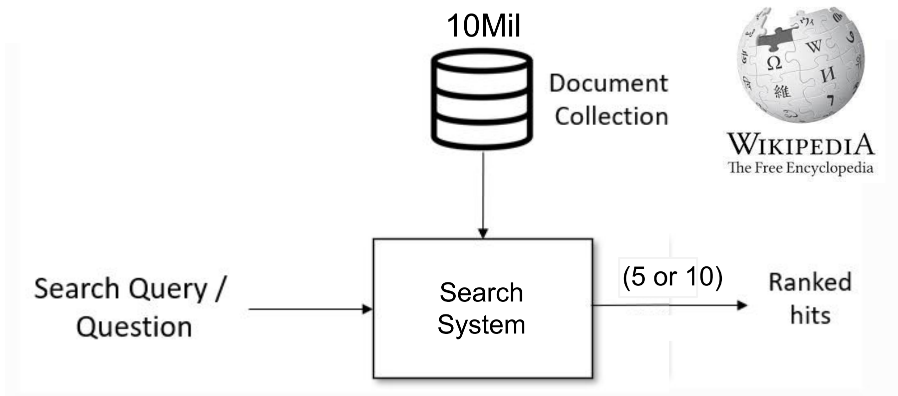
🍻 Why is 🔍 Retrieval Important?

Ubiquitous

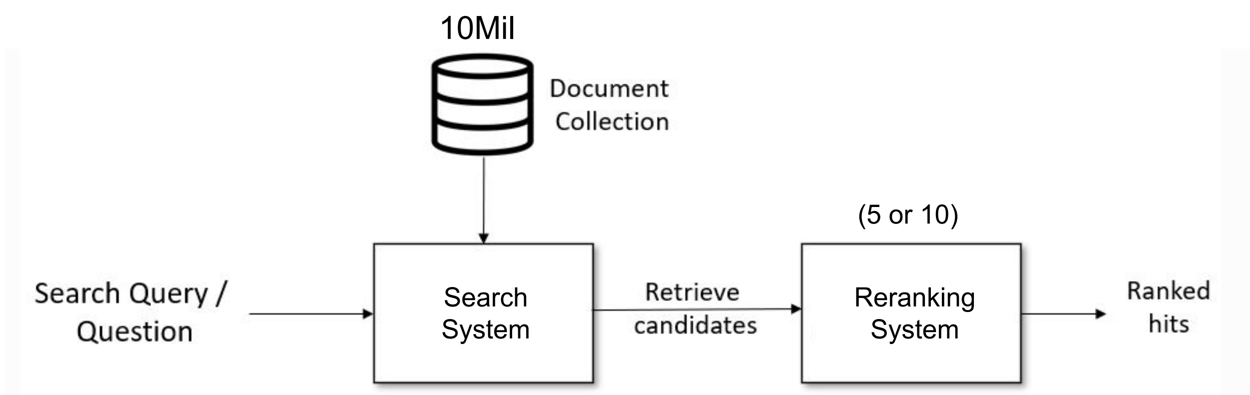
present, appearing, or found everywhere.

Breaking down Retrieval

Information Retrieval



Retrieve and Rerank

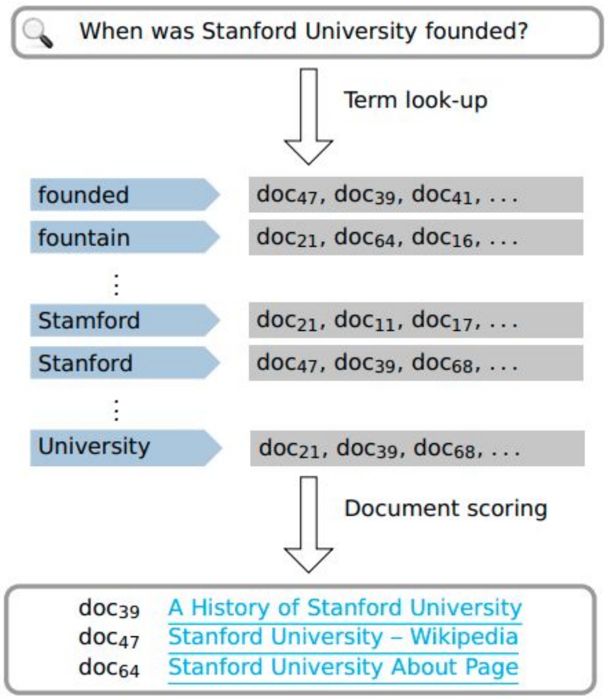


Ref: https://www.sbert.net/examples/applications/retrieve_rerank/README.html

Traditional Search Systems

Traditional Search System: BM25

Keyword-based, Bag of Words Search: Look up keywords from query in documents!



elastic

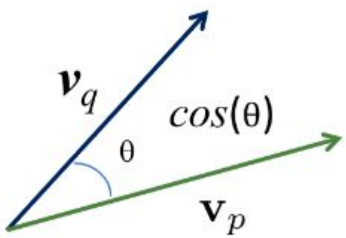
Solr

Lucene

Ref: Christopher G Potts, ACL-IJCNLP 2021 keynote address
https://web.stanford.edu/~cgpotts/talks/potts-acl2021-slides-handout.pdf

Limitations with Traditional Search Systems

Huge Memory Indexes: Sparse vectors are big and can be quite inefficient!



$d_1 >> d_2$

sparse repr: $[0\ldots 1 \ldots. 1 \ldots 0..1] \in \mathbb{R}^{d_1}$

dense repr: $[1.03, -5.72, 6.42, .., 9.91] \in \mathbb{R}^{d_2}$

Unable to handle Synonyms: Won't understand "*bad guy*" and "*villain*" are similar in meaning!

dense

"Who is the **bad guy** in lord of the rings?"

Sala Baker is an actor and stuntman from New Zealand. He is best known for portraying the ***villain*** *Sauron in the Lord of the Rings trilogy by Peter Jackson.*

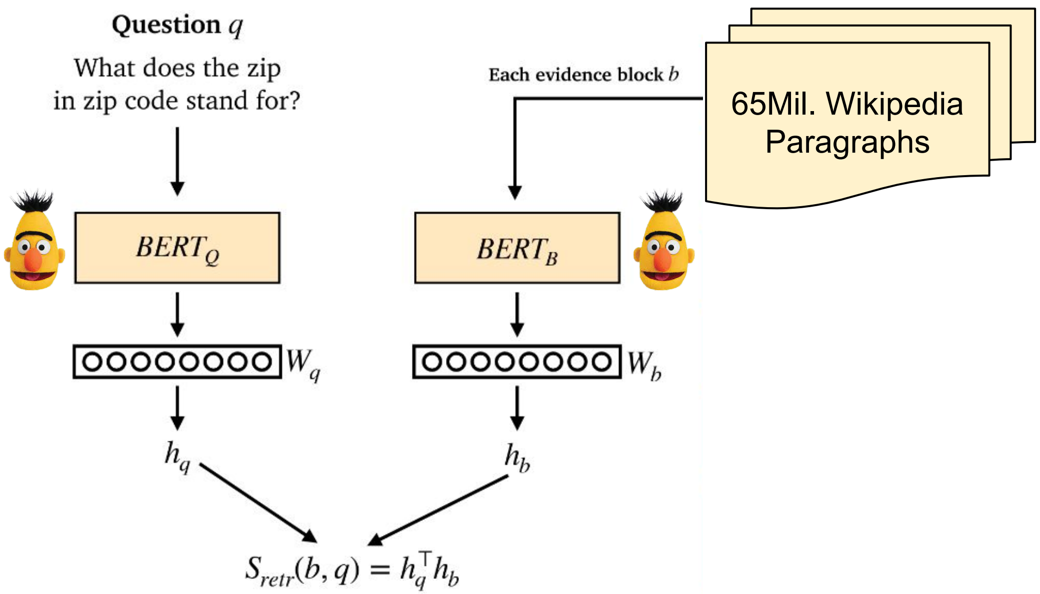
Ref: Danqi Chen, ACL 2020 OpenQA Tutorial
https://github.com/danqi/acl2020-openqa-tutorial/blob/master/slides/part5-dense-retriever-e2e-training.pdf

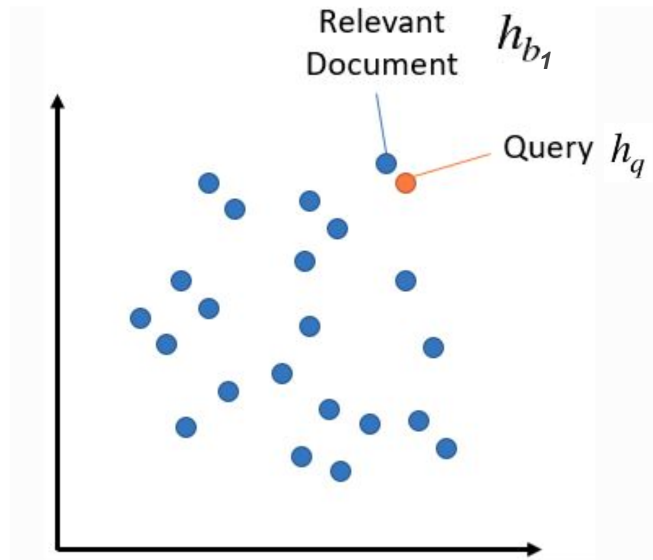
Modern Search Systems

1. **Dense Retrieval: Bi-Encoders**
2. **Reranking: Cross-Encoders**

(1) Dense Retrieval with Bi-Encoders

Bi-Encoders: Encode paragraph and query to a dense vector space with your BERT model !





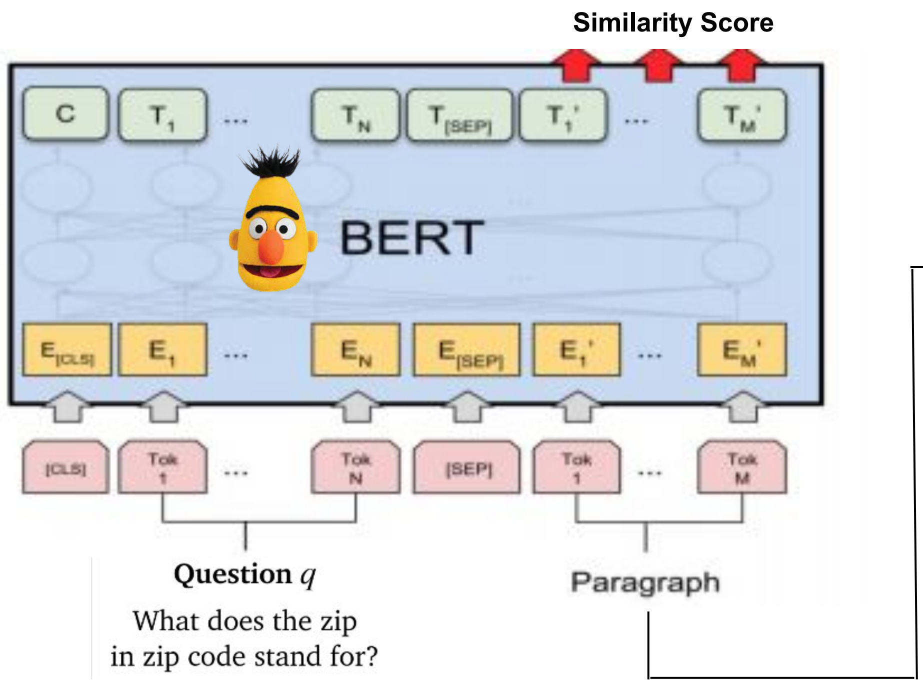
2D Vector Space

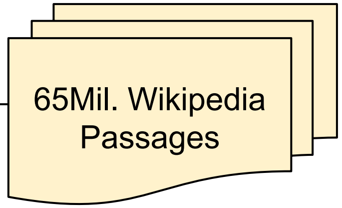
- Passage vectors (hb) can be precomputed and stored!
- Fast and optimal at runtime, ideal for a practical system!

Ref: Danqi Chen, ACL 2020 OpenQA Tutorial
https://github.com/danqi/acl2020-openqa-tutorial/blob/master/slides/part5-dense-retriever-e2e-training.pdf

(2) Reranking with Cross-Encoders

Cross-Encoders: Directly provide paragraph and query to BERT model, No Encoding to vector space !





Scoring thousands or millions of (query, document)-pairs is slow!

Best performance, due to attention across the query and the document.

Ref: Devlin, J., Chang, M-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805v2, 2019.

Traditional vs. Modern Search Systems

Performance: Cross-Encoder **>>** Bi-Encoder **>** BM25

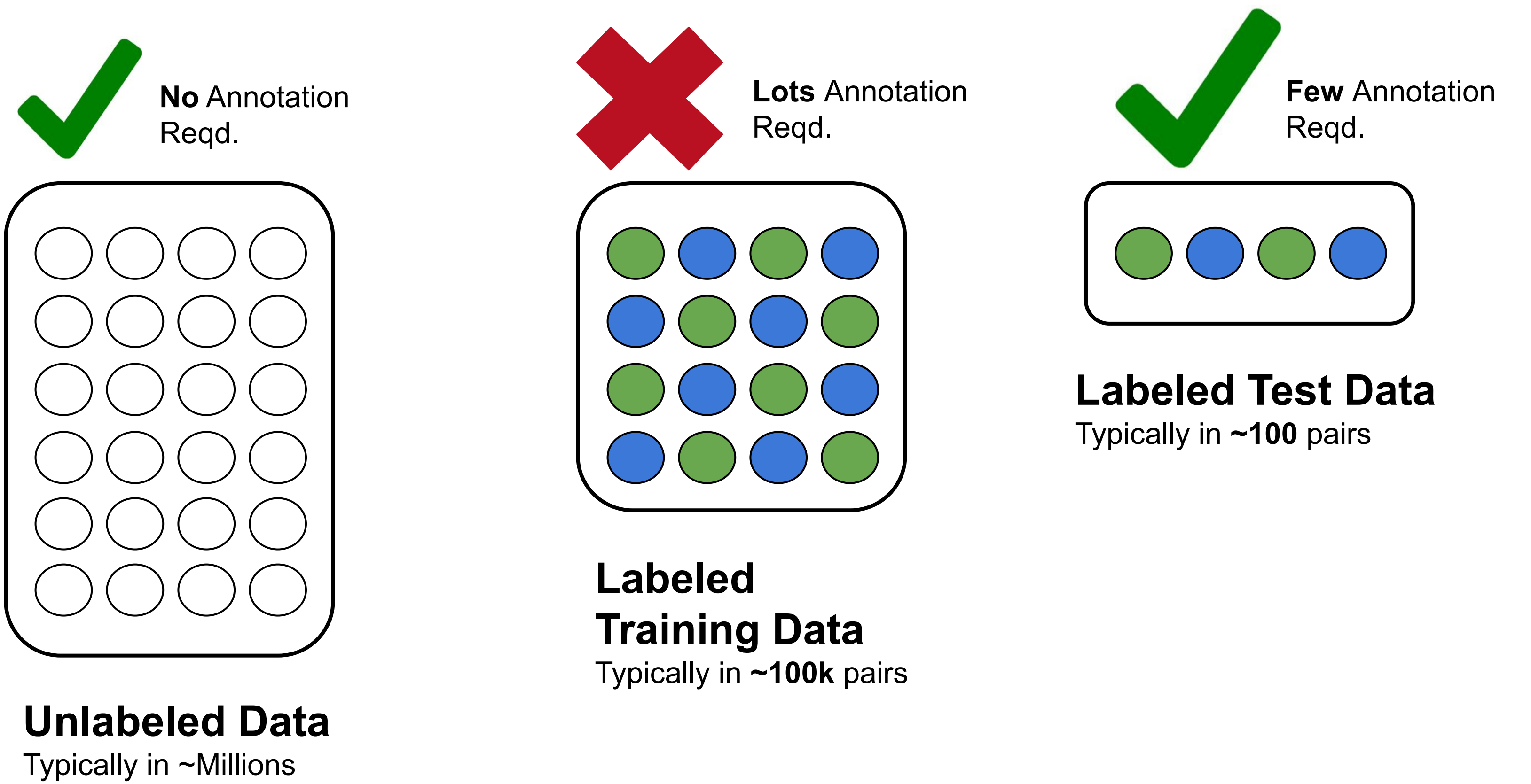
Efficiency: BM25 **>>** Bi-Encoder **>** Cross-Encoder

The Script uses the smaller Simple English Wikipedia as document collection. We test out sample user queries below and compare results:

https://colab.research.google.com/drive/1l6stpYdRMmeDBK_vw0L5NitdiAuhdsAr?usp=sharing

Limitations with Modern Search Systems!

For training/evaluation of **bi-encoders** or **cross-encoder**, you require **three** types of data:



Can Modern Search Systems Generalize?

TECHNISCHE UNIVERSITÄT DARMSTADT

Bi-Encoders / Cross-Encoders

Trained on

WIKIPEDIA
The Free Encyclopedia

Beir
Benchmarking IR

Domains with Training Data

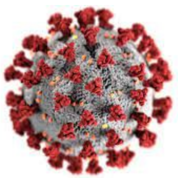
Evaluated on

WIKIPEDIA
The Free Encyclopedia

I can answer any question you have!

Domains without Training Data

Evaluated on



twitter

I'm sorry, I do not understand your question!

BEIR: Evaluation Benchmark for Retrieval Systems

TECHNISCHE UNIVERSITÄT DARMSTADT

Beir

Benchmarking IR

20+ Evaluation datasets!

10+ Diverse domains and tasks!

Make informed decisions on which model to use!

Evaluate and use SOTA models for your own use-case!

Test your own model on our diverse BEIR benchmark!

Easy to use framework, only need to write few lines of code!

BEIR Quick Example

GitHub: https://github.com/UKPLab/beir

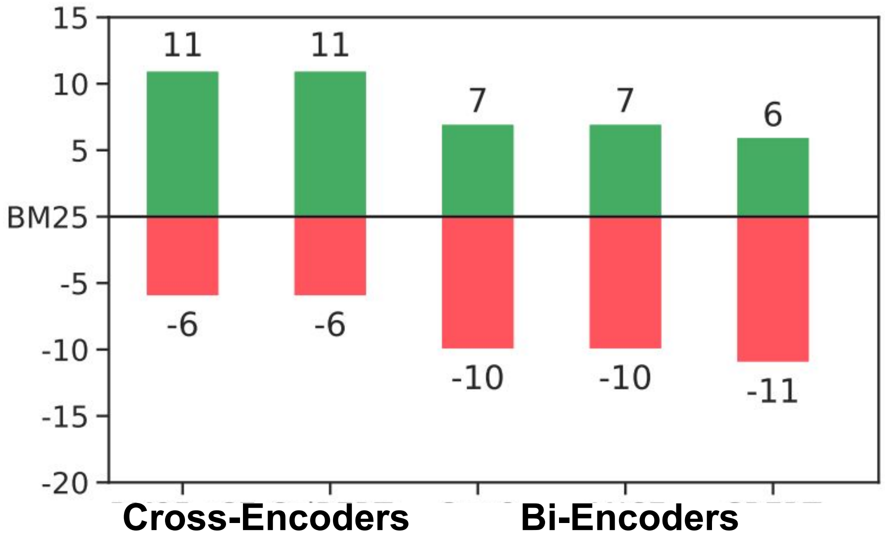
```
### Install BEIR: pip install beir

#### Provide the data_path where scifact has been downloaded and unzipped
corpus, queries, qrels = GenericDataLoader(data_folder=data_path).load(split="test")

#### Load the SBERT model and retrieve using cosine-similarity
model = DRES(models.SentenceBERT("msmarco-distilbert-base-v3"), batch_size=16)
retriever = EvaluateRetrieval(model, score_function="cos_sim") # or "dot" for dot-product
results = retriever.retrieve(corpus, queries)

#### Evaluate your model with NDCG@k, MAP@K, Recall@K and Precision@K  where k = [1,3,5,10,100,1000]
ndcg, _map, recall, precision = retriever.evaluate(qrels, results, retriever.k_values)
```

Steps to Follow:

1. pip install beir
2. Download a BEIR dataset
3. Load BEIR dataset
4. Load Model (Bi-encoder)
5. Evaluate Model on dataset
6. Use model to search on user queries

Ref: Code snippet created using Carbon App. https://carbon.now.sh

Results: Performance Comparison on BEIR



Performance Comparison on BEIR with 17 datasets.

BM25 is the baseline system.

Green denotes modern search systems are better on #datasets. **Red** vice-versa.

BM25 (Lexical)	**Cross-Encoders (Rerank)**	**Bi-Encoders (Dense)**
BM25 is an overall strong system. It doesn't require to be trained.	Reranking Models generalize best. They outperform BM25 on **11/17** retrieval datasets.	Dense models suffer from generalization. They outperform BM25 on **7/17** datasets.

Ref: Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. arXiv preprint arXiv:2104.08663.

🍻Efficiency and Memory Comparison on BEIR

DBPedia (1 Million)			Retrieval Latency		Index
Rank	Model	Dim.	GPU	CPU	Size
(1)	Cross-Encoders	768	550ms	7100ms	0.4GB
(2)	Cross-Encoders	128	350ms	–	20GB
(3)	BM25	–	–	20ms	0.4GB
(4)	Bi-Encoders	768	14ms	125ms	3GB
(5)	Bi-Encoders	768	20ms	275ms	3GB
(6)	Bi-Encoders	768	14ms	125ms	3GB

BM25 (Lexical)

BM25 is overall **fast** and **efficient**. They require small indexes.

Cross-Encoders (Rerank)

Rerankers are **slow** at retrieval. They can also produce **bulky** indexes for retrieval.

Bi-Encoders (Dense)

Dense retrievers are **fast** and **efficient**. They consume less memory with **small** indexes.

Ref: Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. arXiv preprint arXiv:2104.08663.

To Recap

Traditional vs Modern Search Systems

1. Traditional Search Systems like BM25 use keyword based-search which miss out on Synonyms.
2. Bi-Encoders map query and document to a dense vector space, efficient and practical.
3. Cross-Encoders take the query and document together, best performing.
4. Generalization with models is quite a difficult task and there is no free lunch!

How you can use BEIR Benchmark for your own use-case?

5. Let's say you are company working on patent detection.
6. Identify duplicate patents in your system for a new patent which claims to be novel.
7. Use our BEIR Benchmark to see which model is best suited for your task!
8. Cheers! You are happy that you can find duplicate patents with the best model on BEIR!

Thank You For Listening!
Any Questions?

Paper Link: https://arxiv.org/abs/2104.08663

BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Abstract

Neural IR models have often been studied in homogeneous and narrow settings, which has considerably limited insights into their generalization capabilities. To address this, and to allow researchers to more broadly establish the effectiveness of their models, we introduce **BEIR** (***Be**nchmarking **IR***), a *heterogeneous benchmark* for information retrieval.

the keywords also present within the query. Further, queries and documents are treated in a bag-of-words manner which does not take word ordering into consideration.

Recently, deep learning and in particular pre-trained Transformer models like BERT (Devlin et al., 2018) have became popular in the information retrieval space (Lin et al., 2020). They overcome the lexical gap by mapping queries and

GitHub: https://github.com/UKPLab/beir

beir

A Heterogeneous Benchmark for Information Retrieval. Easy to use, evaluate your models across 15+ diverse IR datasets.

Python ★ 213 25

Google colab

https://colab.research.google.com/drive/1HfutiEhHMJLXiWGT8pcipxT5L2TpYEdt?usp=sharing

What is 🔍 Information Retrieval?

🔍 **Longer-term complications of those who recover from COVID-19?**

CORD-19

More than 50 long-term effects of COVID-19: a systematic review and meta-analysis

COVID-19 can involve persistence, sequelae, and other medical complications that last weeks to months after initial recovery. This systematic review and meta-analysis aims to identify studies assessing the long-term effects of COVID-19. LitCOVID and Embase were searched to identify articles with original data published before the 1st of January 2021, with a minimum of 100 patients.

CORD-19

Long term respiratory complications of covid-19

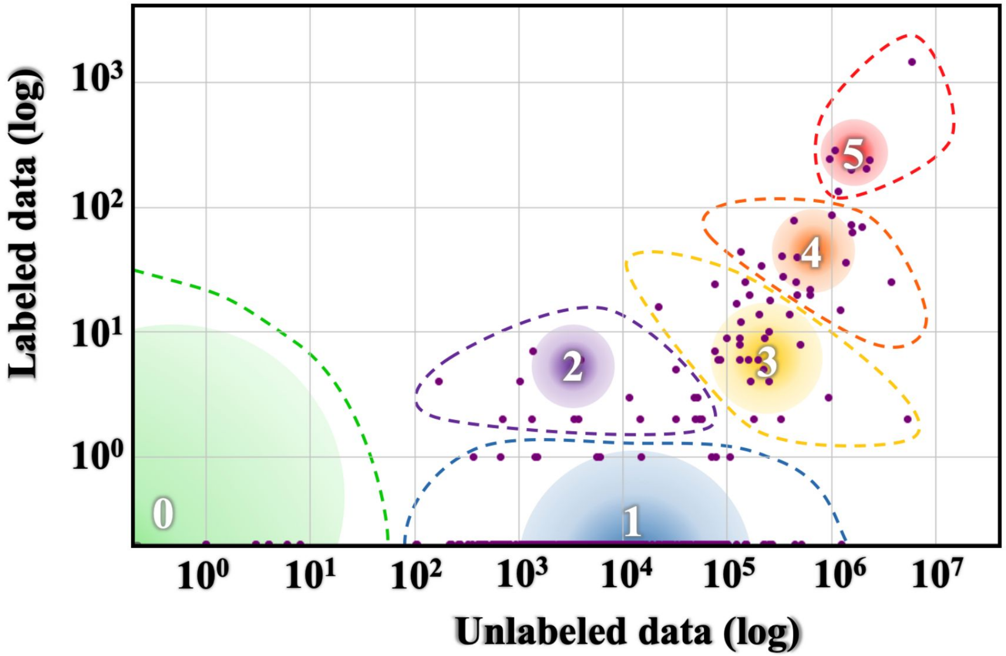
The extent and severity of the long term respiratory complications of covid-19 infection remain to be seen, but emerging data indicate that many patients experience persistent respiratory symptoms months after their initial illness. Recently published guidance by the NHS lays out the likely aftercare needs of patients recovering from covid-19 and identifies potential respiratory problems including chronic cough, fibrotic lung disease, bronchiectasis, and pulmonary vascular disease.

CORD-19

Post–COVID-19 Syndrome: Theoretical Basis, Identification, and Management

As COVID-19 continues to spread, with the United States surpassing 29 million cases, health care workers are beginning to see patients who have been infected with SARS-CoV-2 return seeking treatment for its longer-term physical and mental effects. The term long-haulers is used to identify patients who have not fully recovered from the illness after weeks or months.

Future Work: Datasets in Different Languages



5 - The Winners

English, Spanish, German, Japanese, French

4 - The Underdogs

Russian, Hungarian, Vietnamese, Dutch, Korean

3 - The Rising Stars

Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew

Language resource distribution of **Joshi et al. (2020)**. The size and colour of a circle represent the number of languages and speakers respectively in each category.

Ref: Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. ACL 2020.

Future: Multilingual IR Benchmark (mBEIR)

Language Specific Training Data

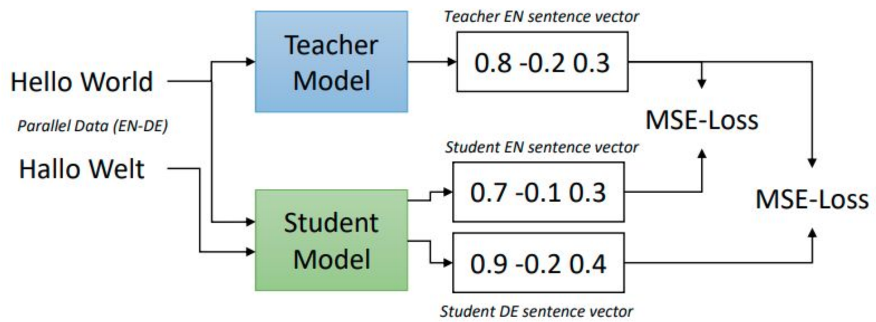
GermanQuAD

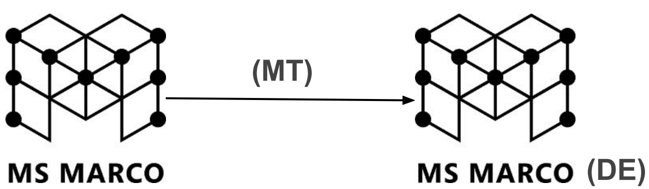
FQuAD

SherQuAD

KorQuAD

Machine-Translated Training Data





Ref: Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. EMNLP 2020.