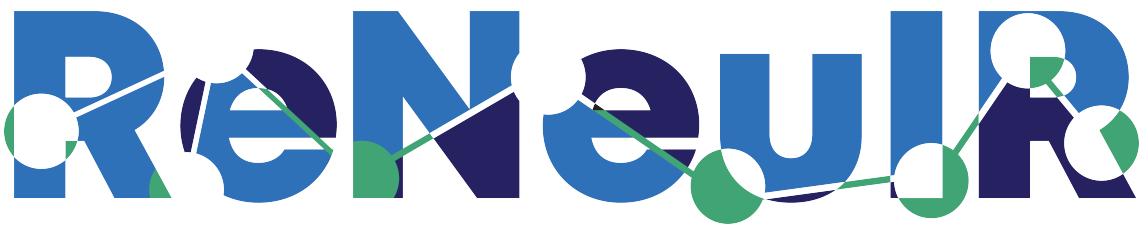
A Proposal for an LSR-Benchmark for Efficiency and Effectiveness-oriented Evaluations



ReNeuIR 2025, July 17, Padua, Italy

Maik Fröbe, Tim Hagen, Matthias Hagen, Heinrich Merker, Franco Maria Nardini, Martin Potthast, Cosimo Rulli, Harry Scells, Ferdinand Schlatt, Rossano Venturini

Additional Collaborators Welcome :)

University of Jena University of Kassel ISTI-CNR
University of Tübingen University of Pisa

A Proposal for an LSR-Benchmark

Learned Sparse Retrieval in a Nutshell

Example document:

Neo lives in a simulation.

A Proposal for an LSR-Benchmark

Learned Sparse Retrieval in a Nutshell

Example document:

Learn how to expand and remove terms

A Proposal for an LSR-Benchmark

Learned Sparse Retrieval in a Nutshell

Example document:

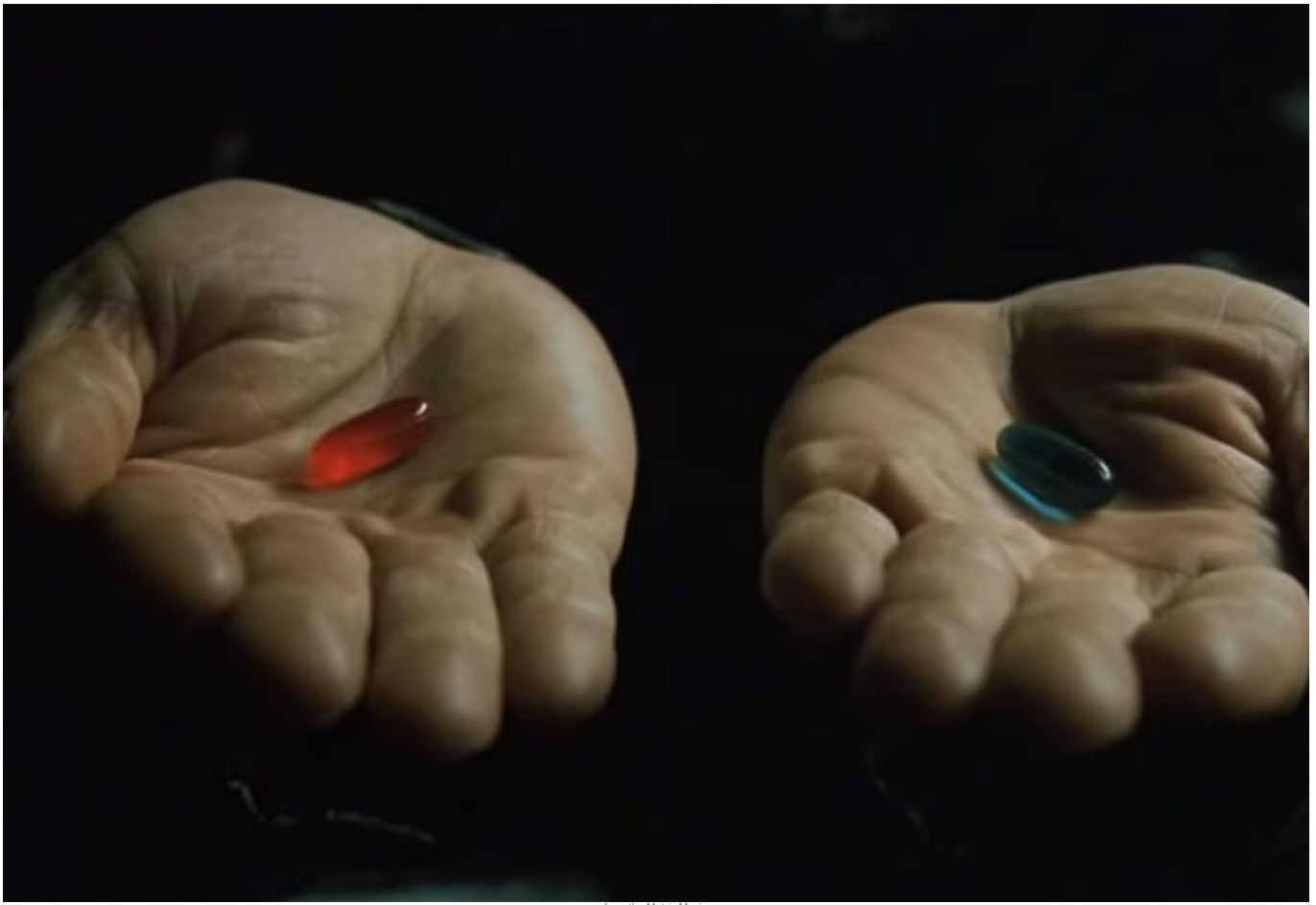
Learn how to expand and remove terms

Interesting trade-offs to study:

- Efficiency: Inverted index can be used
  - But: different term distributions compared to lexical retrieval
- Effectiveness: model learned for specific task
  - Different retrieval tasks might require different term expansion/removal

A Proposal for an LSR-Benchmark

Should we focus our Evaluation on Efficiency or Effectiveness?



from the Matrix Movie

A Proposal for an LSR-Benchmark

Effectiveness-Oriented LSR Experiments

Inputs

- Document texts
- Query texts

Focus

- **Model** to embed documents/queries

A Proposal for an LSR-Benchmark

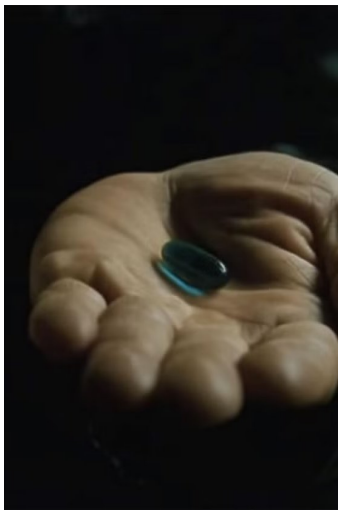
Effectiveness-Oriented LSR Experiments

Inputs

- Document texts
- Query texts

Focus

- **Model** to embed documents/queries



Efficiency-Oriented LSR Experiments

Inputs

- Document embeddings
- Query embeddings

Focus:

- **Index** for approximate retrieval

A Proposal for an LSR-Benchmark

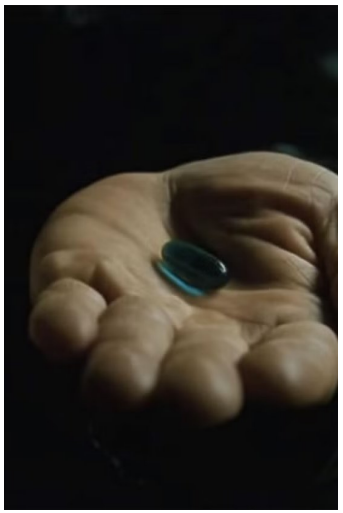
Effectiveness-Oriented LSR Experiments

Inputs

- Document texts
- Query texts

Focus

- **Model** to embed documents/queries



Efficiency-Oriented LSR Experiments

Inputs

- Document embeddings
- Query embeddings

Focus:

- **Index** for approximate retrieval

Goal: Enable Synergies between both for holistic evaluations

A Proposal for an LSR-Benchmark

What Evaluation Corpora Should we use?

- TREC-style evaluation corpora are ideal
  - Reliable evaluations: Pooling for judgments ($>$ 500 judgments per query)
- Still, TREC-style corpora are hard
  - Often very large (e.g., hundreds of GB to a few TB)

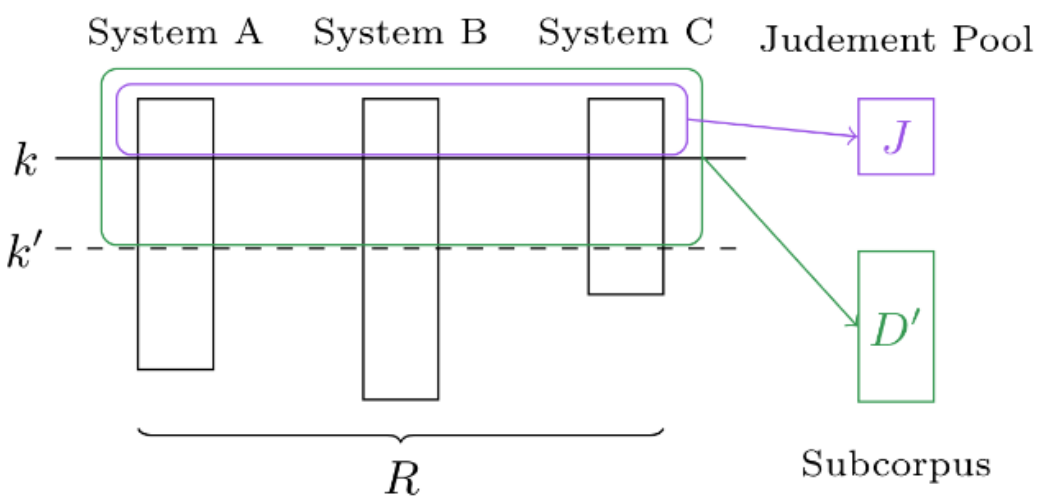
A Proposal for an LSR-Benchmark

What Evaluation Corpora Should we use?

- TREC-style evaluation corpora are ideal
  - Reliable evaluations: Pooling for judgments ($>$ 500 judgments per query)
- Still, TREC-style corpora are hard
  - Often very large (e.g., hundreds of GB to a few TB)

Our Solution: Corpus Subsampling for Fast and Green Evaluation

[ECIR'25]



A Proposal for an LSR-Benchmark

How big are the Corpus Subsamples?

[ECIR'25]

Corpus	Complete			Subsampled		
	Docs.	$\notin_J$	Size	Docs.	$\notin_J$	Size
ClueWeb09	1.0 b	99 %	4.0 TB	0.3 m	73 %	**0.9 GB**
ClueWeb12	0.7 b	99 %	4.5 TB	0.1 m	72 %	**0.5 GB**
MS MARCO	8.8 m	99 %	2.9 GB	0.3 m	97 %	**42.1 MB**

A Proposal for an LSR-Benchmark

How big are the Corpus Subsamples?

[ECIR'25]

Corpus	Complete			Subsampled		
	Docs.	$\notin_J$	Size	Docs.	$\notin_J$	Size
ClueWeb09	1.0 b	99 %	4.0 TB	0.3 m	73 %	**0.9 GB**
ClueWeb12	0.7 b	99 %	4.5 TB	0.1 m	72 %	**0.5 GB**
MS MARCO	8.8 m	99 %	2.9 GB	0.3 m	97 %	**42.1 MB**

Subsampling allows easy sharing/hosting

- Of texts of evaluation corpora for effectiveness-oriented evaluations
- Embeddings for efficiency-oriented evaluations

A Proposal for an LSR-Benchmark

How big are the Corpus Subsamples?

[ECIR'25]

Corpus	Complete			Subsampled		
	Docs.	$\notin_J$	Size	Docs.	$\notin_J$	Size
ClueWeb09	1.0 b	99 %	4.0 TB	0.3 m	73 %	**0.9 GB**
ClueWeb12	0.7 b	99 %	4.5 TB	0.1 m	72 %	**0.5 GB**
MS MARCO	8.8 m	99 %	2.9 GB	0.3 m	97 %	**42.1 MB**

Subsampling allows easy sharing/hosting

- Of texts of evaluation corpora for effectiveness-oriented evaluations
- Embeddings for efficiency-oriented evaluations

Custom Post-Processing

- Re-mapped document IDs to UUIDs to work on custom slices/dices
- Splitting of documents into passages as in MS MARCO v2.1

A Proposal for an LSR-Benchmark

We prepare an ir_datasets like API for frequent use-cases

Step 1: Embedding

```
dataset = lsr_benchmark.load('<IR-DATASETS-ID>')

# process the document texts:
with tracking():
    for doc in dataset.docs_iter(embedding=None):
        doc # namedtuple<doc_id, segments.text>
        # todo: embed document
```

A Proposal for an LSR-Benchmark

We prepare an ir_datasets like API for frequent use-cases

Step 1: Embedding

```
dataset = lsr_benchmark.load('<IR-DATASETS-ID>')

# process the document texts:
with tracking():
    for doc in dataset.docs_iter(embedding=None):
        doc # namedtuple<doc_id, segments.text>
        # todo: embed document
```

Step 2: Indexing and Retrieval

```
dataset = lsr_benchmark.load('<IR-DATASETS-ID>')

# process the document texts:
with tracking():
    for doc in dataset.docs_iter(embedding='<EMBEDDING-MODEL>', passage_aggregation="first-passage"):
        doc # namedtuple<doc_id, embedding>
        # todo: index document
```

A Proposal for an LSR-Benchmark

Conclusions

- Goal: lsr benchmark for holistic evaluations of efficiency/effectiveness
- We use corpus subsampling to enable sharing of corpora and embeddings
- We have a vertical prototype for the 2009 Web Track on the ClueWeb09
- There is much room for improvement and collaboration

Combining Teaching and Research in IR

Conclusions

- Goal: lsr benchmark for holistic evaluations of efficiency/effectiveness
- We use corpus subsampling to enable sharing of corpora and embeddings
- We have a vertical prototype for the 2009 Web Track on the ClueWeb09
- There is much room for improvement and collaboration

Next Steps

- We aim to add one more dataset per week
- ECIR paper of some form

Combining Teaching and Research in IR

Conclusions

- Goal: lsr benchmark for holistic evaluations of efficiency/effectiveness
- We use corpus subsampling to enable sharing of corpora and embeddings
- We have a vertical prototype for the 2009 Web Track on the ClueWeb09
- There is much room for improvement and collaboration

Next Steps

- We aim to add one more dataset per week
- ECIR paper of some form

Thank you!