Corpus Subsampling: Estimating the Effectiveness of Neural Retrieval Models on Large Corpora

ECIR 2025, April 6–10, Lucca, Italy

Maik Fröbe, Andrew Parry, Harrisen Scells, Shuai Wang, Shengyao Zhuang
Guido Zuccon, Martin Potthast and Matthias Hagen

University of Jena University of Glasgow University of Leipzig The University of Queensland

@webis_de www.webis.de

Corpus Subsampling

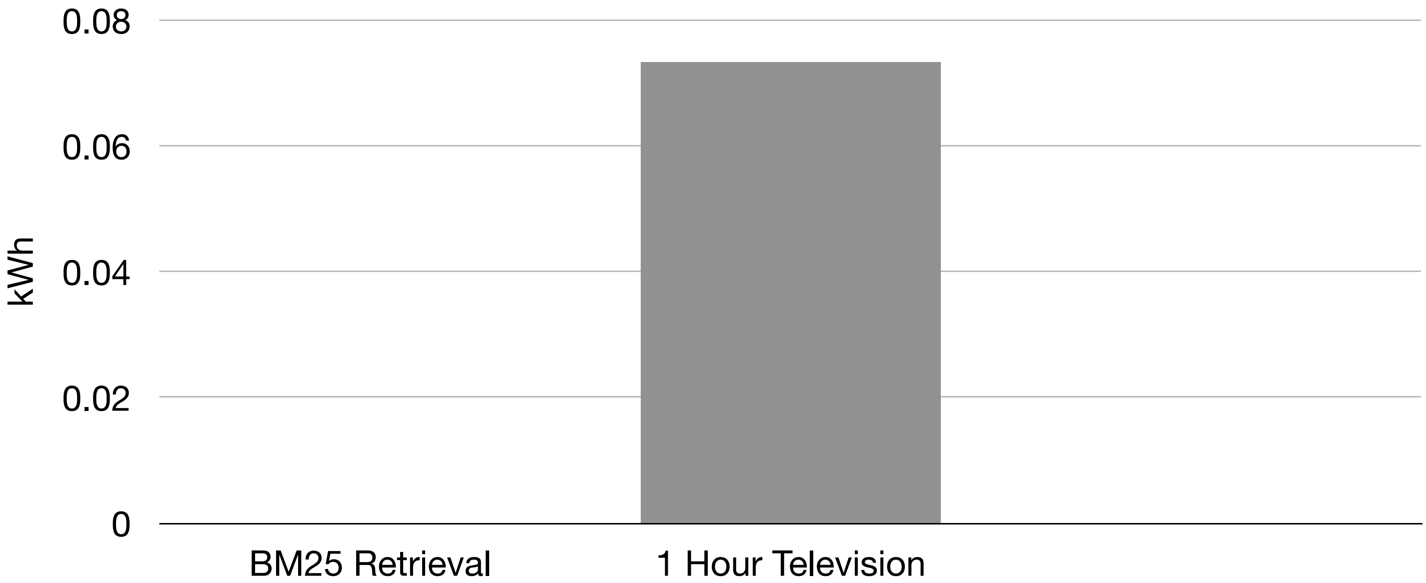
Neural Retrieval Models are Power Hungry

[Scells'22]

Corpus Subsampling

Neural Retrieval Models are Power Hungry

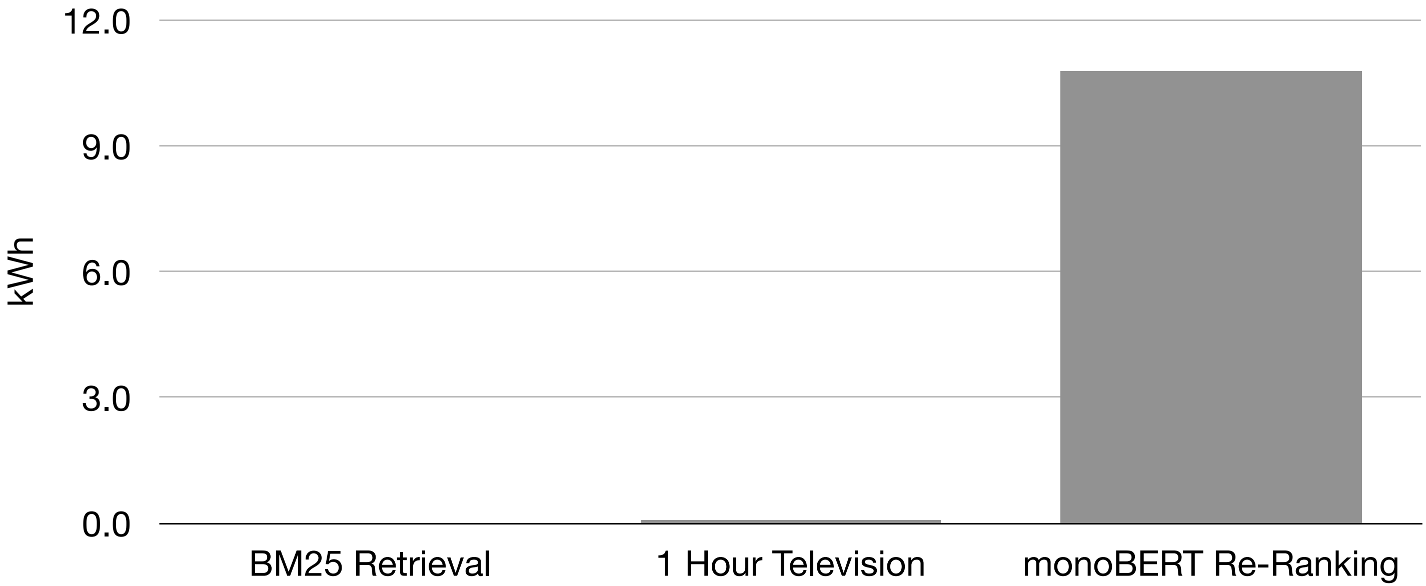
[Scells'22]



Corpus Subsampling

Neural Retrieval Models are Power Hungry

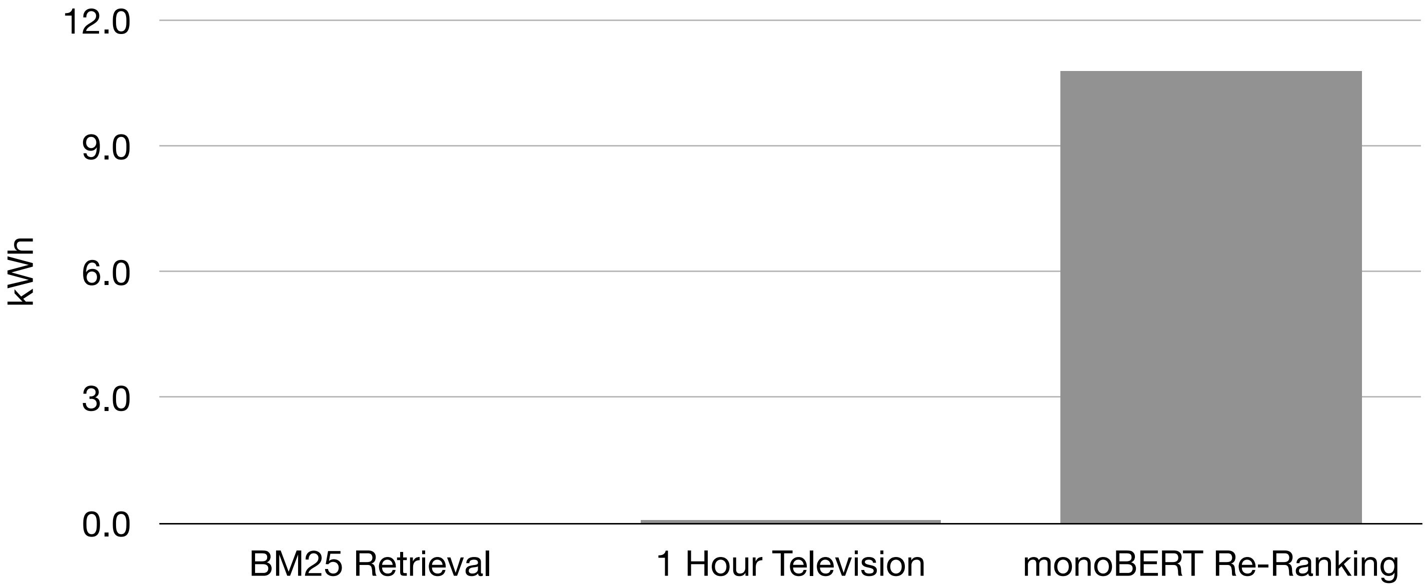
[Scells'22]



Corpus Subsampling

Neural Retrieval Models are Power Hungry

[Scells'22]



Green IR is ...

[Schwartz'20]

Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.

Corpus Subsampling

Considerations to Make Research-Oriented Evaluations Greener

Our Evaluation will always give us some number

- Is this number meaningful?

Corpus Subsampling

Considerations to Make Research-Oriented Evaluations Greener

Our Evaluation will always give us some number

- Is this number meaningful?

Solution: Ensure that our evaluation is reliable
[Voorhees'19]

- Observations transfer to similar scenarios with a high probability

System A > System B

Corpus Subsampling

Considerations to Make Research-Oriented Evaluations Greener

Our Evaluation will always give us some number

- Is this number meaningful?

Solution: Ensure that our evaluation is reliable
[Voorhees'19]

- Observations transfer to similar scenarios with a high probability

System A > System B

Correlations of system rankings can confirm the reliability of evaluations
[Breuer'20]

Step 1: Create a system ranking with all data

System A > Sytem B > System C > System D

Corpus Subsampling

Considerations to Make Research-Oriented Evaluations Greener

Our Evaluation will always give us some number

- Is this number meaningful?

Solution: Ensure that our evaluation is reliable
[Voorhees'19]

- Observations transfer to similar scenarios with a high probability

System A > System B

Correlations of system rankings can confirm the reliability of evaluations
[Breuer'20]

Step 1: Create a system ranking with all data

System A > Sytem B > System C > System D

Step 2: Repeat the experiment in a „greener" setting

New System Ranking	$\tau_{\mathbf{AP}}$	kWh
System A > Sytem B > System C > System D	1.0	1000
System A > Sytem B > System D > System C	0.8	**1**

Corpus Subsampling

How build our Evaluation Dataset? Step 1: Queries

Many queries with few judgments or few queries with many judgments?

Corpus Subsampling

How build our Evaluation Dataset? Step 1: Queries

Many queries with few judgments or few queries with many judgments?

How many different rankings?

Labels				**Top-10 Rankings**
0	1	2	3	
∞	1	—	—	11
∞	10	10	10	$4^{10} >$ 1 million

Corpus Subsampling

How build our Evaluation Dataset? Step 1: Queries

Many queries with few judgments or **few queries with many judgments**?

How many different rankings?

Pooling advantageous
from Green IR Perspective

Labels				**Top-10 Rankings**
0	1	2	3	
∞	1	—	—	11
∞	10	10	10	$4^{10} >$ 1 million

Corpus Subsampling

How build our Evaluation Dataset? Step 2: Documents

Evaluation Corpora with top-k pooling typically:

- Have **50 queries**
- Pool **30 to 100 systems**
- Between **10 million and 1 billion documents**

Corpus Subsampling

How build our Evaluation Dataset? Step 2: Documents

Evaluation Corpora with top-k pooling typically:

- Have **50 queries**
- Pool **30 to 100 systems**
- Between **10 million and 1 billion documents**

What documents to include to evaluate on ca. 50 pooled queries?

Considerations:

- A few million document suffice to satisfy most information needs [Mei'08]
- We do not need to include all relevant documents
- We only need a subset that allows reliable evaluations

Corpus Subsampling

Document Selection Strategies

Judgment Pool:

- Select all documents with a judgment. E.g., the top-10 pool
- Disadvantage: Effectiveness overestimated in post-hoc experiments
  [Sakai'08,Fröbe'23]

Corpus Subsampling

Document Selection Strategies

Judgment Pool:

- Select all documents with a judgment. E.g., the top-10 pool
- Disadvantage: Effectiveness overestimated in post-hoc experiments [Sakai'08,Fröbe'23]

Re-Ranking:

- Select all documents retrieved by a model. E.g., the top-1k of BM25
- Disadvantage: Bias towards the first stage model

Corpus Subsampling

Document Selection Strategies

Judgment Pool:

- Select all documents with a judgment. E.g., the top-10 pool
- Disadvantage: Effectiveness overestimated in post-hoc experiments [Sakai'08,Fröbe'23]

Re-Ranking:

- Select all documents retrieved by a model. E.g., the top-1k of BM25
- Disadvantage: Bias towards the first stage model

Judgment Pool + Random

- All documents with a judgment plus random documents
- Disadvantage: Random documents are too easy negatives

Corpus Subsampling

Document Selection Strategies

Judgment Pool:

- Select all documents with a judgment. E.g., the top-10 pool
- Disadvantage: Effectiveness overestimated in post-hoc experiments
  [Sakai'08,Fröbe'23]

Re-Ranking:

- Select all documents retrieved by a model. E.g., the top-1k of BM25
- Disadvantage: Bias towards the first stage model

Judgment Pool + Random

- All documents with a judgment plus random documents
- Disadvantage: Random documents are too easy negatives

Re-Pooling

- Re-Pool to $k^{'} >> k$. E.g., top-100 or 1k for a top-10 judgment pool
- Advantage: Incorporates many distractors. Can use all above.

Corpus Subsampling

Evaluation: Reliability of System Rankings

Experiments on 9 evaluation campaigns on four corpora

- ClueWeb09, ClueWeb12, Robust04, MS MARCO

Corpus Subsampling

Evaluation: Reliability of System Rankings

Experiments on 9 evaluation campaigns on four corpora

- ClueWeb09, ClueWeb12, Robust04, MS MARCO

Leave-one-Group-out Experiments

- For each team, assume all systems of the team did not participate
- Remove documents only retrieved by the team from the judgments/corpus
- Re-Evaluate all systems and compare their ranking with the ground truth

Corpus Subsampling

Evaluation: Reliability of System Rankings

Experiments on 9 evaluation campaigns on four corpora

- ClueWeb09, ClueWeb12, Robust04, MS MARCO

Leave-one-Group-out Experiments

- For each team, assume all systems of the team did not participate
- Remove documents only retrieved by the team from the judgments/corpus
- Re-Evaluate all systems and compare their ranking with the ground truth

Results

Subsampling	τ_{PJ}			
	ClueWeb09	**ClueWeb12**	**Robust04**	**MS MARCO**
Judgment Pool	0.944	0.941	0.983	0.978
Re-Ranking BM25	0.936	0.938	0.836	0.994
Judgment Pool + Random	0.799	0.765	0.789	0.794
Re-Pooling $k^{'} = 100$	**0.980**	**0.987**	**0.995**	**0.999**

Corpus Subsampling

Evaluation: Reliability of System Rankings

Subsampling modifies the corpus statistics

- Unretrieved Documents can impact the ranking of the top documents
- Ranking on a subsample should mimick retrieval from the complete corpus

Corpus Subsampling

Evaluation: Reliability of System Rankings

Subsampling modifies the corpus statistics

- Unretrieved Documents can impact the ranking of the top documents
- Ranking on a subsample should mimick retrieval from the complete corpus

Experimental setup

- 9 evaluation campaigns (ClueWeb09, ClueWeb12, Robust04, MS MARCO)
- 10 lexical models, 7 Bi-Encoder models, 3 Late Interaction models
- RBO correlation against retrieval from all retrieved documents

Corpus Subsampling

Evaluation: Reliability of System Rankings

Subsampling modifies the corpus statistics

- Unretrieved Documents can impact the ranking of the top documents
- Ranking on a subsample should mimick retrieval from the complete corpus

Experimental setup

- 9 evaluation campaigns (ClueWeb09, ClueWeb12, Robust04, MS MARCO)
- 10 lexical models, 7 Bi-Encoder models, 3 Late Interaction models
- RBO correlation against retrieval from all retrieved documents

Results

Subsampling	**ClueWeb09**		
	Bi-E.	Late	Lex.
Judgment Pool	.297	.263	.295
Re-Ranking BM25	.139	.192	.037
Judgment Pool + Random	.096	.111	.056
Re-Pooling $k' = 100$	**.600**	**.481**	**.660**

Corpus Subsampling

How big are the resulting subcorpora?

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
Disks 4/5	0.5 m	41 %	0.6 GB	0.4 m	31 %	0.5 GB
MS MARCO	8.8 m	99 %	2.9 GB	0.3 m	97 %	42.1 MB

Corpus Subsampling

Conclusions

- Pooling can produce subcorpora for reliable post-hoc evaluation
- Allows to evaluate expensive retrieval approaches on large corpora
- Subsamples improve accessibility:
  - E.g., a reliable ClueWeb09 subsample is 0.9 GB
- Leave-one-group-out simulations to the reliability of a subsample in advance

Corpus Subsampling

Conclusions

- Pooling can produce subcorpora for reliable post-hoc evaluation
- Allows to evaluate expensive retrieval approaches on large corpora
- Subsamples improve accessibility:
  - E.g., a reliable ClueWeb09 subsample is 0.9 GB
- Leave-one-group-out simulations to the reliability of a subsample in advance

Future Work

- Can corpus subsampling be integrated into evaluation campaigns?
  - Step 1: Run evaluation campaign on huge, noisy corpora
  - Step 2: Subsample corpus
  - All post-hoc experiments run on the subsample

Corpus Subsampling

Conclusions

- Pooling can produce subcorpora for reliable post-hoc evaluation
- Allows to evaluate expensive retrieval approaches on large corpora
- Subsamples improve accessibility:
  - E.g., a reliable ClueWeb09 subsample is 0.9 GB
- Leave-one-group-out simulations to the reliability of a subsample in advance

Future Work

- Can corpus subsampling be integrated into evaluation campaigns?
  - Step 1: Run evaluation campaign on huge, noisy corpora
  - Step 2: Subsample corpus
  - All post-hoc experiments run on the subsample

Thank you!

Results (2)

Subsampling	$\Delta_{\text{nDCG@10}}$			
	ClueWeb09	**ClueWeb12**	**Robust04**	**MS MARCO**
Judgment Pool	0.030	0.031	0.005	0.011
Re-Ranking BM25	-0.013	-0.053	0.049	-0.005
Judgment Pool + Random	0.375	0.325	0.062	0.259
Re-Pooling $k' = 100$	-0.030	-0.060	-0.004	-0.007