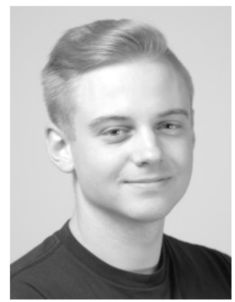
Team Galápagos Tortoise at LongEval 2024: Neural Re-Ranking and Rank Fusion for Temporal Stability

September 12, 2024

Marlene Gründel

Malte Weber



Johannes Franke

Jan Heinrich Merker

Friedrich-Schiller-Universität Jena

https://webis.de

15th International Conference of the CLEF Association (CLEF 2024)

Team Galápagos Tortoise at LongEval 2024

Motivation

- Modern retrieval systems use multi-stage re-ranking
- Static test collections prone to train-test leakage
  - → Unrealistic scenario
- Models struggle with temporal changes

→ Let's develop systems that maintain effectiveness over time

(How Stable Diffusion thinks BERT would explain multi-stage re-ranking.)

Team Galápagos Tortoise at LongEval 2024

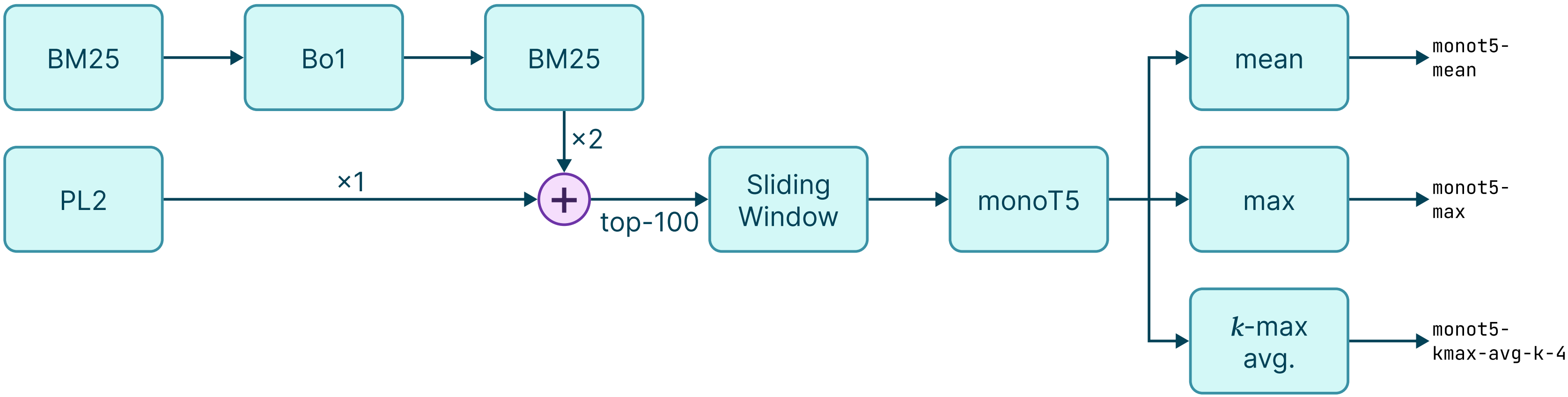
Our Research

1. Explore passage score aggregations for monoT5 re-ranking
   - ❑ Standard bi-encoder re-ranking after lexical first-stage retrieval
   - ❑ Passage score aggregations: max, mean, k-max average
2. Explore rank fusion of diverse retrieval models
   - ❑ LLM-based re-ranker after lexical first-stage retrieval
   - ❑ Fusion with cross-encoder, late-interaction, and lexical

➔ Evaluate effectiveness and temporal stability
   - ❑ nDCG
   - ❑ Decline over time: Jan$\rightarrow$Jun, Jun$\rightarrow$Aug, Jan$\rightarrow$Aug

Team Galápagos Tortoise at LongEval 2024

Approach: monoT5 Re-Ranking



- Initial retrieval: Weighted combination of BM25 and PL2
- Re-ranking top-50 results with monoT5
- Comparing passage score aggregation schemes:
  - Max passage
  - Mean passage
  - k-max average ($k = 4$, tuned on Jan. data)

Team Galápagos Tortoise at LongEval 2024

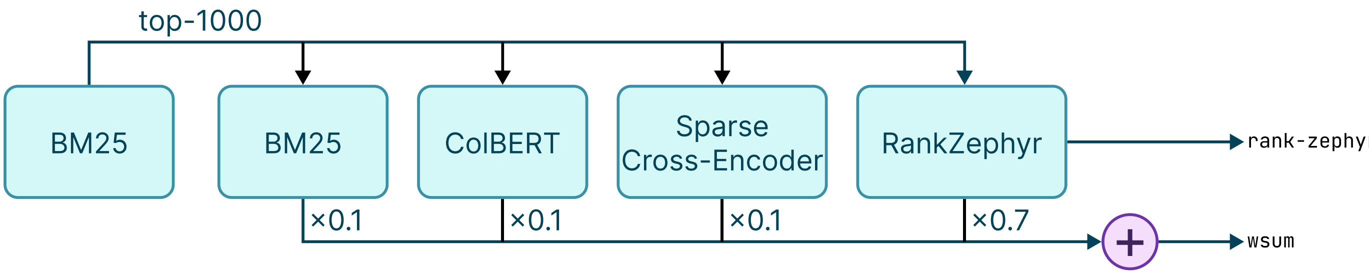
Results: monoT5 Re-Ranking

- Max passage aggregation outperforms mean passage
- Difference more significant on recent datasets
- k-max average worse than max passage
- All systems show temporal decline in effectiveness

System	nDCG@10		nDCG	
	value	p value	value	p value
January 2023				
max passage	**0.209**	—	**0.307**	—
4-max avg. passage	0.208	0.86	0.305	0.41
mean passage	0.209	0.93	0.307	0.82
June 2023				
max passage	**0.196**	—	**0.260**	—
4-max avg. passage	0.191	0.24	0.257	0.24
mean passage	0.184	0.02	0.253	0.02
August 2023				
max passage	**0.159**	—	**0.198**	—
4-max avg. passage	0.156	0.07	0.196	0.12
mean passage	0.150	$<$0.01	0.191	$<$0.01

Team Galápagos Tortoise at LongEval 2024

Approach: Rank Fusion



- Weighted rank fusion of:
  - RankZephyr (weight: 0.7)
  - Sparse Cross-Encoder (weight: 0.1)
  - ColBERT (weight: 0.1)
  - BM25 (weight: 0.1)
- Optimized for nDCG@10 on January 2023 dataset

Team Galápagos Tortoise at LongEval 2024

Results: Rank Fusion

- Rank fusion significantly outperforms most individual models
- No significant difference between fusion and just RankZephyr
- Highly effective systems (fusion, RankZephyr) show greater temporal decline

System name	nDCG@10		nDCG	
	value	p value	value	p value
January 2023				
rank fusion	**0.251**	—	**0.355**	—
RankZephyr	0.247	0.07	0.353	0.26
Sparse Cross-Enc.	0.221	<0.01	0.337	<0.01
ColBERT	0.216	<0.01	0.330	<0.01
June 2023				
rank fusion	0.228	—	0.293	—
RankZephyr	**0.228**	0.98	**0.295**	0.34
Sparse Cross-Enc.	0.202	$<$0.01	0.277	$<$0.01
ColBERT	0.183	$<$0.01	0.264	$<$0.01
August 2023				
rank fusion	**0.180**	—	**0.220**	—
RankZephyr	0.178	0.15	0.219	0.52
Sparse Cross-Enc.	0.169	$<$0.01	0.212	$<$0.01
ColBERT	0.161	$<$0.01	0.206	$<$0.01

Team Galápagos Tortoise at LongEval 2024

Summary

- Max passage aggregation most effective/stable for monoT5 re-ranking
- Rank fusion improves effectiveness but not temporal stability
- All systems, including BM25, show effectiveness decline over time
- Future work:
  - Investigate temporal decline in lexical models
  - Explore more fusion candidates

Code and Data

github.com/webis-de/CLEF-24

Summary

- Max passage aggregation most effective/stable for monoT5 re-ranking
- Rank fusion improves effectiveness but not temporal stability
- All systems, including BM25, show effectiveness decline over time
- Future work:
  - Investigate temporal decline in lexical models
  - Explore more fusion candidates

Code and Data

github.com/webis-de/CLEF-24

Thank you & merci!