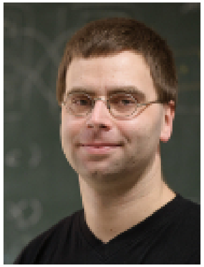
Noise-Reduction for Automatically Transferred Relevance Judgments



Maik Fröbe[1] Christopher Akiki[2] Martin Potthast[2] Matthias Hagen[1]

Martin-Luther-Universität Halle-Wittenberg[1] Leipzig University[2]

CLEF, 5–8 September 2022

webis.de

Noise-Reduction for Automatically Transferred Relevance Judgments

Motivation: Transition from Version 1 to Version 2 of MS MARCO

Version 1:

- Document dataset crawled in 2018
- URL matching to transfer relevance judgments from the passage-level
- Gap between judgments and crawling: 1 year

Motivation: Transition from Version 1 to Version 2 of MS MARCO

Version 1:

- Document dataset crawled in 2018
- URL matching to transfer relevance judgments from the passage-level
- Gap between judgments and crawling: 1 year

Version 2:

- Document dataset crawled in 2021
- Larger and cleaner (improves encoding, passage-document mapping, etc.)
- Relevance judgments transferred from Version 1 with URL matching
- Gap between judgments and crawling: 4 years

Noise-Reduction for Automatically Transferred Relevance Judgments

Motivation: Transition from Version 1 to Version 2 of MS MARCO

Version 1:

- Document dataset crawled in 2018
- URL matching to transfer relevance judgments from the passage-level
- Gap between judgments and crawling: 1 year

Version 2:

- Document dataset crawled in 2021
- Larger and cleaner (improves encoding, passage-document mapping, etc.)
- Relevance judgments transferred from Version 1 with URL matching
- Gap between judgments and crawling: 4 years

Observation in the 2021 DL Track:

[Craswell et all, TREC’21 Notebooks]

Models trained on Version 1 more effective than models trained on Version 2

Contributing Factors to the Effectiveness Drop

Content of positive training documents might have changed

- Document in Version 1 is relevant to its query
- Document in Version 2 is not relevant to its query

Contributing Factors to the Effectiveness Drop

Content of positive training documents might have changed

- Document in Version 1 is relevant to its query
- Document in Version 2 is not relevant to its query

Example:

Query	Relevant Document	
	Version 1 (2018)	Version 2 (2021)
what are yellow roses mean	Meaning Of A Yellow Rose … a yellow rose *stands for joy and happiness* …	20 Best Knockout Roses To Make Your Garden Outstanding

Contributing Factors to the Effectiveness Drop

Content of positive training documents might have changed

- Document in Version 1 is relevant to its query
- Document in Version 2 is not relevant to its query

Example:

Query	**Relevant Document**	
	Version 1 (2018)	Version 2 (2021)
what are yellow roses mean	Meaning Of A Yellow Rose . . . a yellow rose *stands for joy and happiness* . . .	20 Best Knockout Roses To Make Your Garden Outstanding

Goal:

Assess prevalence of such noise in the training data

Noise-Reduction for Automatically Transferred Relevance Judgments

MonoT5 to Identify of Candidates for Noisy Training Instances

- Trained on the passage dataset of MS MARCO
- Max-Passage aggregation
- Estimates $P(\text{Relevant} = 1|d, q)$

Noise-Reduction for Automatically Transferred Relevance Judgments

MonoT5 to Identify of Candidates for Noisy Training Instances

- Trained on the passage dataset of MS MARCO
- Max-Passage aggregation
- Estimates $P(\text{Relevant} = 1|d, q)$

	Relevant Document	
	Version 1 (2018)	Version 2 (2021)
Query: what are yellow roses mean	Meaning Of A Yellow Rose … a yellow rose *stands for joy and happiness* …	20 Best Knockout Roses To Make Your Garden Outstanding
$P(\text{Rel} = 1 \mid d, q)$	0.92	0.04

Noise-Reduction for Automatically Transferred Relevance Judgments

MonoT5 to Identify of Candidates for Noisy Training Instances

- ❑ Trained on the passage dataset of MS MARCO
- ❑ Max-Passage aggregation
- ❑ Estimates $P(\text{Relevant} = 1|d, q)$

	Relevant Document	
	Version 1 (2018)	Version 2 (2021)
Query: what are yellow roses mean	Meaning Of A Yellow Rose . . . a yellow rose *stands for joy and happiness* . . .	20 Best Knockout Roses To Make Your Garden Outstanding
$P(\text{Rel} = 1|d, q)$	0.92	0.04

Filter-Criteria for Error Candidates in Version 2

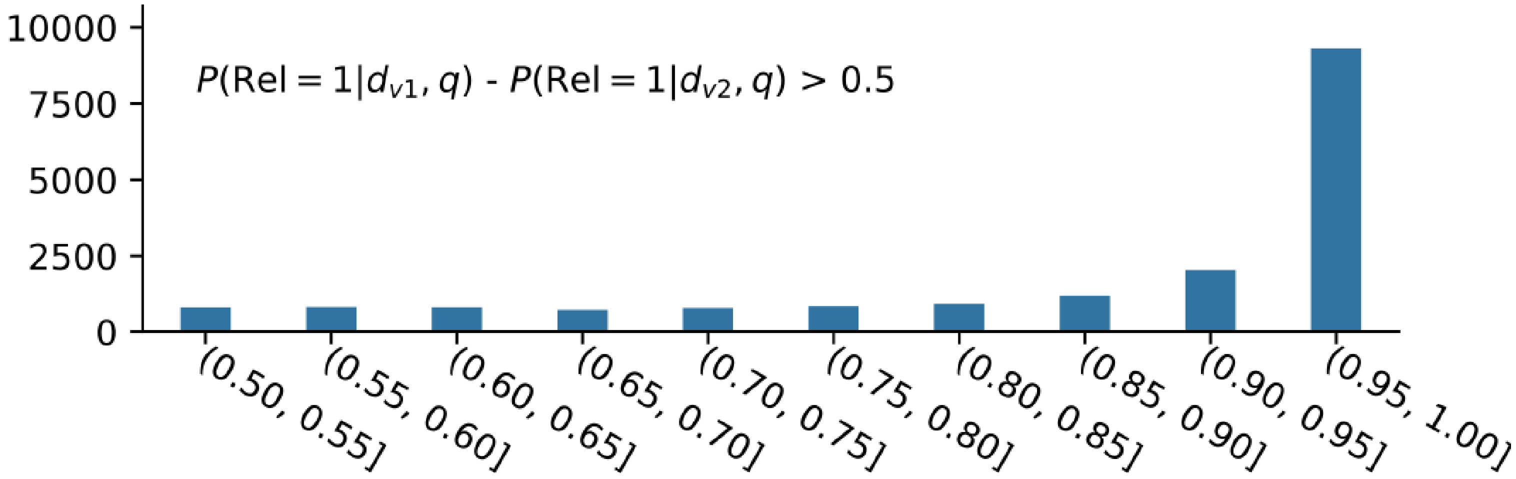
MonoT5 estimates document in Version 1 substantially more relevant:

$$P(\text{Relevant} = 1|d_{v1}, q) - P(\text{Relevant} = 1|d_{v2}, q) > 0.5$$

Noise-Reduction for Automatically Transferred Relevance Judgments

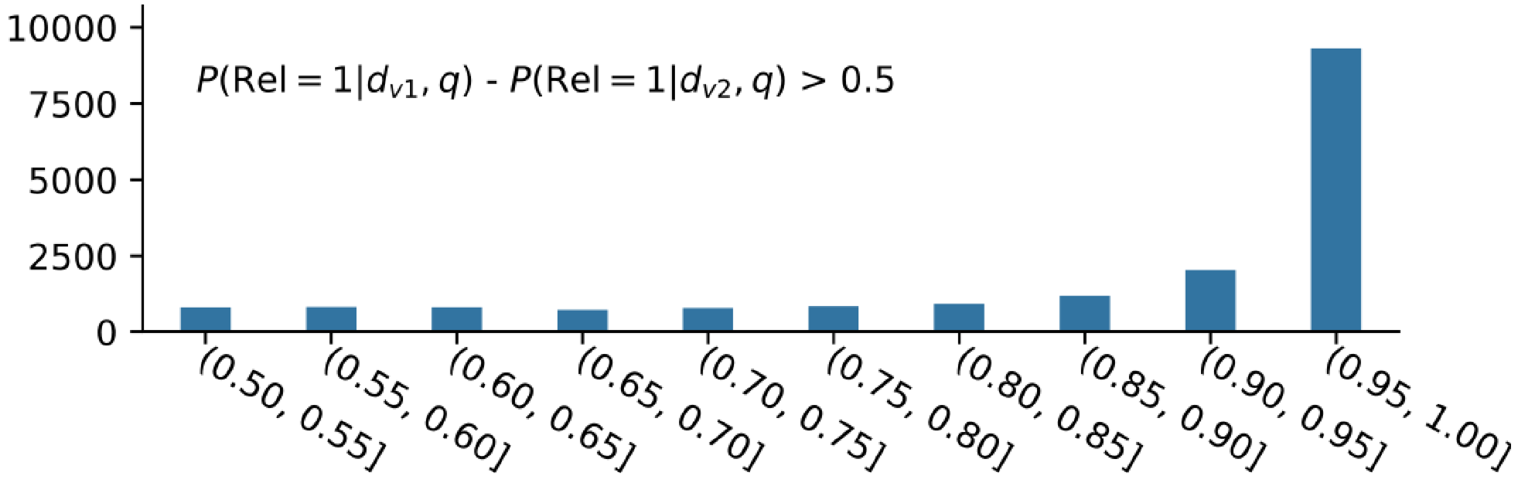
Error Candidates in Version 2 identified by MonoT5 (1)

17,969 error candidates



Error Candidates in Version 2 identified by MonoT5 (1)

17,969 error candidates



Manual Verification of Error Candidates:

- Review of 100 random error candidates
- Precision: 0.73
- Estimated number or errors: $17,969 \cdot 0.73 = \mathbf{13,117}$

Error Candidates in Version 2 identified by MonoT5 (2)

Do error candidates negatively affect the effectiveness of trained models?

- We train monoT5-base models on queries from the error candidates
- Training queries have error candidates in Version 2
- Repeat experiments 10 times with varying seeds

Retr. Model		**nDCG@10**		
Model	Version	ClueWeb12	DL 19/20	Robust04
BM25	—	0.298	0.507	0.449
monoT5	1	**0.387**†	**0.562**†	**0.446**†
monoT5	2	0.177	0.142	0.209

Reverse Direction: Error Candidates in Version 1

MonoT5 estimates document in Version 2 substantially more relevant:

Query	**Relevant Document**	
	Version 1 (2018)	Version 2 (2021)
what are deposit solutions banking	Oops! There was a problem! We had an unexpected problem processing your request.	*Deposit Solutions* Crunchbase *Company Profile* . . .

Reverse Direction: Error Candidates in Version 1

MonoT5 estimates document in Version 2 substantially more relevant:

Query	**Relevant Document**	
	Version 1 (2018)	Version 2 (2021)
what are deposit solutions banking	Oops! There was a problem! We had an unexpected problem processing your request.	*Deposit Solutions* Crunchbase *Company Profile* ...

We find 15,817 error candidates

- Precision in manual review of 100 random candidates: 0.25
- Estimated number or errors: $15,817 \cdot 0.25 = \mathbf{3,954}$

Reverse Direction: Error Candidates in Version 1

MonoT5 estimates document in Version 2 substantially more relevant:

Query	**Relevant Document**	
	Version 1 (2018)	Version 2 (2021)
what are deposit solutions banking	Oops! There was a problem! We had an unexpected problem processing your request.	*Deposit Solutions* Crunchbase *Company Profile* . . .

We find 15,817 error candidates

- Precision in manual review of 100 random candidates: 0.25
- Estimated number or errors: $15,817 \cdot 0.25 = \mathbf{3,954}$

Do error candidates negatively affect the effectiveness of trained models?

Retr. Model		**nDCG@10**		
Model	Version	ClueWeb12	DL 19/20	Robust04
monoT5	1	0.238	0.316	0.279
monoT5	2	**0.318**†	**0.476**†	**0.367**†

Takeaways

Comparison of monoT5 scores across both versions of MS MARCO

- Positive document from Version 1 or Version 2?
- Retrieval models trained on the “wrong” version are highly ineffective

Using Version 2 of MS MARCO for training is discouraged now
[Craswell et all, TREC’21]

- Models learn to prioritize “old” content
- Support from our experiments:
  - 3,954 estimated errors in Version 1 vs. 13,117 in Version 2

Takeaways

Comparison of monoT5 scores across both versions of MS MARCO

- Positive document from Version 1 or Version 2?
- Retrieval models trained on the “wrong” version are highly ineffective

Using Version 2 of MS MARCO for training is discouraged now
[Craswell et all, TREC’21]

- Models learn to prioritize “old” content
- Support from our experiments:
  - 3,954 estimated errors in Version 1 vs. 13,117 in Version 2

Future work:

- Other factors: Preprocessing pipelines, etc.
- More fine-grained classification of content changes

Code, Paper, Slides: webis.de/publications

Noise-Reduction for Automatically Transferred Relevance Judgments

Takeaways

Comparison of monoT5 scores across both versions of MS MARCO

- Positive document from Version 1 or Version 2?
- Retrieval models trained on the “wrong” version are highly ineffective

Using Version 2 of MS MARCO for training is discouraged now
[Craswell et all, TREC’21]

- Models learn to prioritize “old” content
- Support from our experiments:
  - 3,954 estimated errors in Version 1 vs. 13,117 in Version 2

Future work:

- Other factors: Preprocessing pipelines, etc.
- More fine-grained classification of content changes

Code, Paper, Slides: webis.de/publications

Thank You!