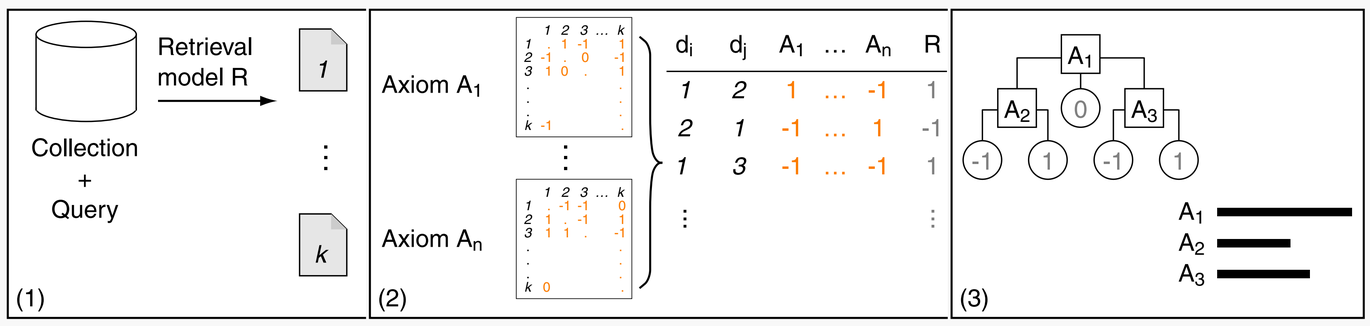
Michael Völske[1] Alexander Bondarenko Maik Fröbe
Benno Stein Jaspreet Singh Matthias Hagen Avishek Anand

[1]michael.voelske@uni-weimar.de

Towards Axiomatic Explanations for Neural Ranking Models

Our Axiomatic Explanation Framework Explains Rankings by Reconstruction



Top: Our axiomatic explanation pipeline. (1) The retrieval model to be explained ranks a document set of size *k*. (2) Axioms produce ranking preferences for all pairs. (3) A simple explanation model reconstructs the initial ranking.

Right: Two examples of axioms, and the twenty axioms included in our explanation framework, including three different implementations each for STMC1 and STMC2.

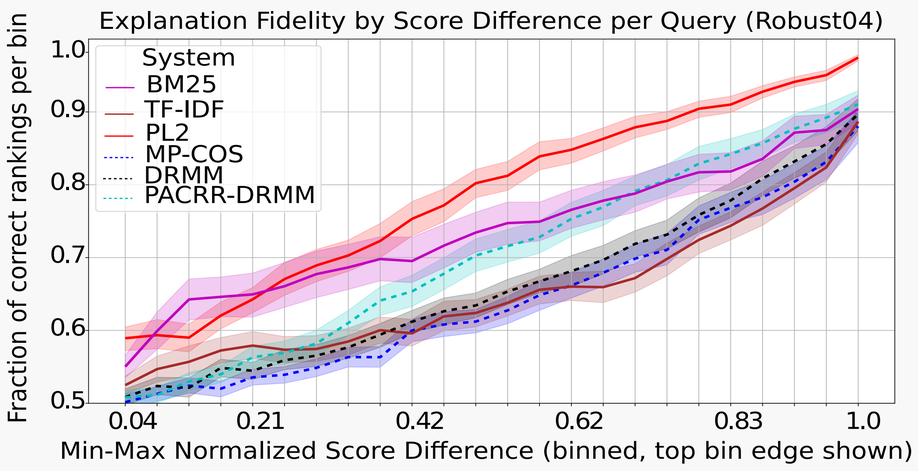
Term frequency: TFC1 [Fang, 2004], TFC3 [Fang, 2011], TDC [Fang, 2004]
Document length: LNC1 [Fang, 2004], TF-LNC [Fang, 2004]
Lower-bounding term frequency: LB1 [Lv, 2011]
Query aspects: REG [Wu, 2012], AND [Zheng, 2010], DIV [Gollapudi, 2009]
Semantic Similarity: STMC1* [Fang, 2006], STMC2* [Fang, 2006]
Term proximity: PROX1–PROX5 [Hagen, 2016]

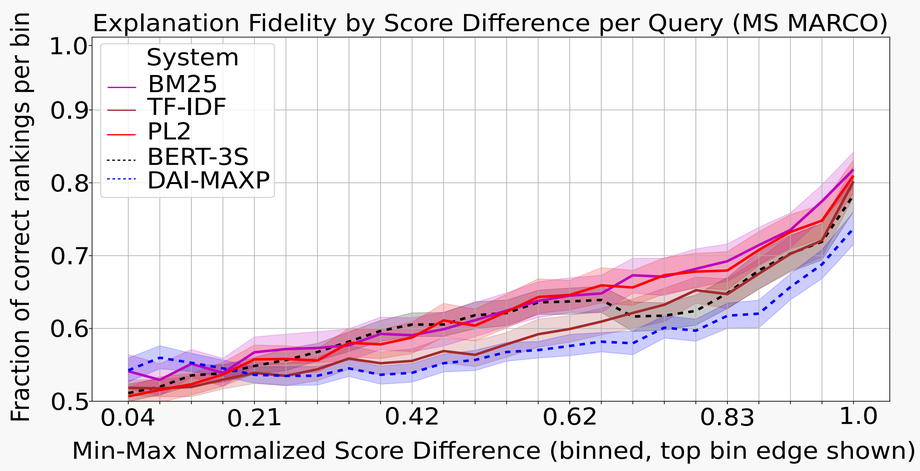
With our Current Axiom Set, Explanation Fidelity is Limited Except for Distant Pairs

Explanation Models		Explanation Fidelity (Fraction Correct Pairs)					
Scope	Per Retr. Model	Classical Retrieval Models			Neural Retrieval Models		
Robust04		**BM25**	**TF-IDF**	**PL2**	**MP-COS**	**DRMM**	**PACRR-DRMM**
query	100	**0.75**	**0.66**	0.78	**0.67**	**0.68**	**0.72**
rank-diff bin	24	0.71	0.63	0.77	0.59	0.61	0.67
score-diff bin	24	0.72	0.64	0.78	0.59	0.61	0.68
query, rank-diff bin	2,368	0.73	0.64	0.77	0.65	0.66	0.70
query, score-diff bin	2,394	0.74	0.65	**0.79**	0.64	0.66	0.70
MS MARCO		**BM25**	**TF-IDF**	**PL2**		**BERT-3S**	**DAI-MAXP**
query	100	**0.64**	**0.60**	**0.63**		**0.61**	**0.59**
rank-diff bin	24	0.60	0.56	0.59		0.57	0.54
score-diff bin	24	0.61	0.56	0.59		0.59	0.55
query, rank-diff bin	2,400	0.62	0.58	0.61		0.60	0.57
query, score-diff bin	2,376	0.63	0.60	0.62		0.61	0.58

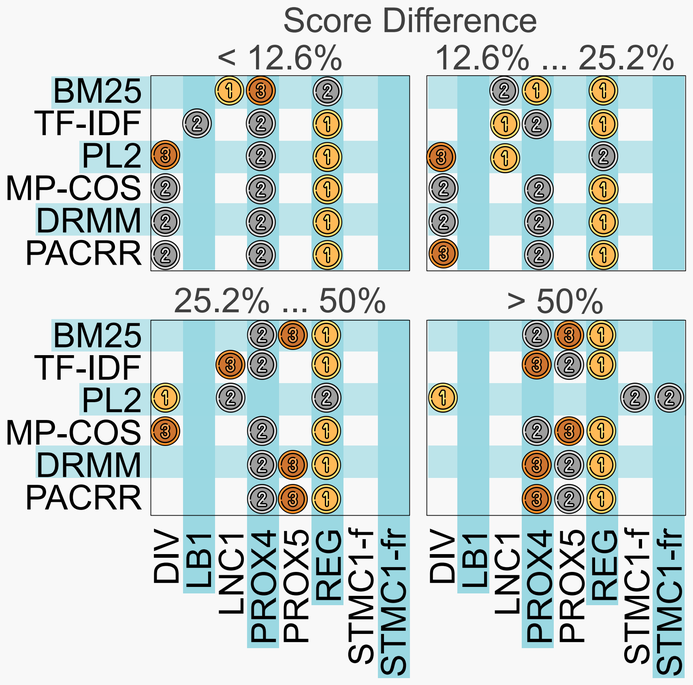
Left: Explanation models for 100 queries each from Robust04 and MS MARCO. For the score- and rank-difference-scope models, document pair samples are divided into 24 bins based on the min-max normalized difference in retrieval scores. Explanation fidelity is computed for each bin separately and is further macro-averaged.

Below: The increase in explainability as the score difference grows is more pronounced for Robust04, whereas the explanations on both datasets perform very similarly at the low-score difference end (0.5 is no better than random guessing).





Top Axioms Overlap Between Models, and "Easier" Queries are Easier to Explain



Query	Explanation fidelity			Topic title	nDCG		
	DRMM	MP-COS	PACRF		DRMM	MP-COS	PACRR
344	0.57	0.55	0.60	Abuses of E-Mail	0.26	0.27	0.33
352	0.64	0.56	0.61	British Chunnel impact	0.15	0.17	0.15
...							
618	0.82	0.81	0.83	Ayatollah Khomeini death	0.41	0.52	0.44
684	0.67	0.55	0.66	Part-time benefits	0.41	0.26	0.39

Top: Explanation fidelity and nDCG are weakly positively correlated (Pearson ≈ 0.1).

Left: Top-3 axioms per retrieval model by relative score difference (Robust04).

Conclusions:

1. Twenty well understood IR axioms to explain black-box neural rankers.
2. Explanation fidelity on the smaller, more genre-focused Robust04 with its shorter queries is superior to that on MS MARCO.
3. Well-grounded axiomatic constraints capturing other retrieval aspects seem to be needed to further improve explanation fidelity.