Towards Axiomatic Explanations for Neural Ranking Models

Michael Völske,[*] Alexander Bondarenko,[†] Maik Fröbe,[†] Benno Stein,[*]
Jaspreet Singh,[‡] Matthias Hagen,[†] Avishek Anand[§]
[*]Bauhaus-Universität Weimar [†]Martin-Luther-Universität Halle-Wittenberg [‡]Amazon [§]Leibniz Universität Hannover

ABSTRACT

Recently, neural networks have been successfully employed to improve upon state-of-the-art effectiveness in ad-hoc retrieval tasks via machine-learned ranking functions. While neural retrieval models grow in complexity and impact, little is understood about their correspondence with well-studied IR principles. Recent work on interpretability in machine learning has provided tools and techniques to understand neural models in general, yet there has been little progress towards explaining ranking models.

We investigate whether one can explain the behavior of neural ranking models in terms of their congruence with well understood principles of document ranking by using established theories from axiomatic IR. Axiomatic analysis of information retrieval models has formalized a set of constraints on ranking decisions that reasonable retrieval models should fulfill. We operationalize this axiomatic thinking to reproduce rankings based on combinations of elementary constraints. This allows us to investigate to what extent the ranking decisions of neural rankers can be explained in terms of the existing retrieval axioms, and which axioms apply in which situations. Our experimental study considers a comprehensive set of axioms over several representative neural rankers. While the existing axioms can already explain the particularly confident ranking decisions rather well, future work should extend the axiom set to also cover the other still "unexplainable" neural IR rank decisions.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**.

KEYWORDS

Axiomatic IR; Explanation; Reproducing Rankings; Neural Models

ACM Reference Format:
Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, Avishek Anand. 2021. Towards Axiomatic Explanations for Neural Ranking Models. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3471158.3472256

ICTIR '21, July 11, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8611-1/21/07…$15.00
https://doi.org/10.1145/3471158.3472256

1 INTRODUCTION

When using machine learning models to rank search results, the training data (clicks, human annotations) drives the way features are combined as relevance signals by various models, ranging from linear regression and decision trees to deep neural networks more recently [37, 41, 46]. On a more abstract level, by learning how to combine features to best rank documents, a machine-learned model indirectly encodes the query intent. Documents are then ordered by relevance, i.e., how well they match the underlying query intent. While such models have achieved state-of-the-art effectiveness in ad-hoc text retrieval, their complexity makes them difficult to interpret, up to the point that eventually some of the reported performance gains have been called into question [28].

Axiomatic thinking on relevance scoring functions has arisen from a similar concern regarding a lack of rigor in formulating what makes a good result ranking. Empirically, non-optimal parameter settings had been shown to cause existing retrieval models to perform poorly, resulting in heavy parameter tuning. Axiomatic practitioners have hence formalized desirable, elementary properties of ranking functions. The analysis of popular scoring functions with respect to their adherence to axiomatic constraints has given rise to revised scoring functions with provably superior performance [30], re-ranking approaches that improve an existing ranking's adherence [23], as well as datasets for diagnosing retrieval models' axiom adherence empirically [7, 44].

In this paper, we apply established retrieval axioms to ground the behavior of arbitrary ranking models in a well understood (and hence interpretable) axiomatic basis. The central question that we raise in this paper is: "To what extent can we explain neural models in terms of the existing IR axioms?" To this end, we follow Hagen et al. [23] and operationalize the retrieval axioms as Boolean predicates that, given a pair of documents and a query, express a preference for either one document or the other to be ranked higher. Given multiple axioms, this yields a set of possibly conflicting ranking preferences for each document pair. Thus, to explain a given result ranking, we compute axiom preferences across all of its constituent document pairs and then fit an explanation model, which is here a simple classification model trained to make the same pairwise ordering decisions as the initial ranking, given only the axiom preferences as predictors. This approach permits various insights into the model that produced the initial ranking: (a) the explanation model's parameters reveal the degree to which different axiomatic constraints are important to the retrieval model under consideration, and (b) the fidelity with which the initial ranking can be reconstructed can point to blind spots in the axiom set, which can help to uncover new ranking properties yet to be formalized.

Several previous studies have used retrieval axioms to explain or improve ranking decisions, but our work is the first to combine many axioms together to specifically try to reconstruct neural rankings in an extensive experimental study. Our paper makes the following contributions: (1) We propose a general-purpose framework to analyze arbitrary rankings with respect to their adherence to information retrieval axioms. (2) In an extensive experimental study on the Robust04 [52] and MS MARCO [40] test collections, we investigate to what extent five different state-of-the-art neural retrieval models can be explained under the axiomatic framework.[1] (3) We explore notions of locality in axiomatic explanations, such as whether different explanations apply to different queries, or to different locations in result rankings.

Our results show that the degree to which rankings can be explained with the currently known axioms is still rather limited overall. The axiomatic explainability of neural rankers tends to be on par with that of simpler classical retrieval models. Large differences in the retrieval score, where retrieval functions are highly confident of the relevance differences they indicate, are well explainable with axioms, but this is much less the case for small fluctuations among closer result documents. Interesting directions for future work thus are the formulation of additional axioms that capture other angles of relevance and that apply in a wide range of real-world contexts, as well as the development of a more rigorous approach to relaxing the preconditions of the existing axioms for a wider applicability.

2 RELATED WORK

The fundamental task of information retrieval—extracting from a large collection those information items relevant to a particular information need—is typically accomplished by ranking the items in a collection according to assigned relevance scores. Most commonly-used scoring functions such as BM25 have been designed to quantify some specific, narrow notion of relevance [2, 16]. Unlike recently-popular machine learning models, axiomatic IR aims to explain a "good" retrieval function by means of mathematically described formal constraints. Possibly the earliest ideas that are close to "axiomatic" IR were a retrieval system complemented with the production rules from artificial intelligence [33], which improved the performance of a Boolean model, a formalization of a conditional logic underlying information retrieval [51], and terminological logic to model IR processes [35]. The first real mention of the term axiom in relation to IR was introduced in a study by Bruza and Huibers [5], who proposed to describe retrieval mechanisms by axioms expressed in terms of concepts from the information field.

In the last decades, the number of studies developing new axioms that describe what a good retrieval function looks like, has considerably increased. More than 20 distinct axioms have been proposed so far, which can be divided into groups by the particular aspect of the relevance scoring problem that they aim to formalize: term frequency [14–16, 39] and lower bounds on it [30, 31], document length [10, 14], query aspects [20, 53, 60], semantic similarity [13, 17], term proximity [23, 50], axioms for evaluation [3, 6], axioms describing properties implied by link graphs [1], axioms for learned ranking functions [8, 9], multi-criteria relevance scoring [19], user-rating based ranking [58], translation language model axioms [25, 43], and term dependency [12]. The majority of the aforementioned studies consider the axioms individually.

Table 1: The 20 retrieval axioms included in our study; STMC1 and STMC2 each implemented in three variants (*).

Purpose	Axioms / Sources
Term frequency	TFC1 [14], TFC3 [15], TDC [14]
Document length	LNC1 [14], TF-LNC [14]
Lower-bounding term frequency	LB1 [30]
Query aspects	REG [53], AND [60], DIV [20]
Semantic Similarity	STMC1* [17], STMC2* [17]
Term proximity	PROX1–PROX5 [23]

The first large-scale study on axioms' impact on retrieval effectiveness was published by Hagen et al. [23], who combined 23 individual axioms to re-rank top-k result sets and showed that different axiom combinations significantly improve the retrieval performance of basic models such as BM25, Terrier DPH, or DirichletLM. We follow this approach of operationalizing retrieval axioms as computable functions, which enables the exploitation of a large set of axioms. However, instead of manipulating rankings to more closely match axiomatic preferences, we reconstruct a retrieval model's rankings via weighted axiom preferences to better understand and explain the decisions underlying the ranking function.

Recently, retrieval axioms have been used to regularize neural retrieval models to prevent over-parameterization, improving both training time and retrieval effectiveness [45]. Retrieval axioms have also been applied to weight meta-learner features which predict how to combine the relevance scores of different retrieval models into an overall score [4]. Rennings et al. [44] have developed a pipeline to create diagnostic datasets for neural retrieval models, each fulfilling one axiom, which allows to detect what kind of axiomatically expressed search heuristics neural models are able to learn. The study's diagnostic datasets focus on only 4 simple individual axioms, which cannot completely account for neural rankers' decisions. In a follow-up publication, Câmara and Hauff [7] extend the idea to building diagnostic datasets for 9 axioms separately, with a focus on BERT-based rankers. MacAvaney et al. [32] systematize the analysis of neural IR models as a framework comprising three testing strategies—controlled manipulation of individual measurements (e.g., term frequency or document length), manipulating document texts, and constructing tests from non-IR datasets—whose influence on neural rankers' behavior can be investigated. In our study, we follow a fourth approach: reconstructing rankings based on elementary axiomatic properties. In the process, we combine the ideas of 20 retrieval axioms at the same time (cf. Table 1; variants of STMC1 and STMC2 are described in Section 3.1), aiming to capture and explain arbitrary neural rankers' decisions. While the aforementioned studies on axiomatic diagnostics of neural rankers typically operate in very controlled and synthetic settings, our approach provides a complementary view based on more realistic, TREC-style queries and document collections.

[1] For the sake of reproducibility, the code for the experiments described in this paper is made available at https://github.com/webis-de/ICTIR-21

The rationales behind the decisions of complex learning systems are also studied in the field of interpretability in machine learning [26, 59]. Recent work on extracting feature attributions using post-hoc approximations is similar to our model of explaining document preference pairs [18, 47–49]. However, we crucially differ from them in two ways: we use axioms as possible explanations and we employ a learning framework to measure the fidelity to the original model rather than a combinatorial framework [47].

3 AXIOMATICALLY EXPLAINING RANKINGS

Our axiomatic explanation framework generates post-hoc explanations for the ranked result lists produced by the retrieval model under investigation. The two main components are a set of *axioms*, and an *explanation model*. Figure 1 provides a high-level overview of the framework: Given a ranked list of documents $(d_1, \ldots, d_k)$ and a set of axioms $\{A_1, \ldots, A_n\}$, we first compute the pairwise ranking preferences for every document pair under each axiom. Thus, each document pair (d_i, d_j) is associated with an n-dimensional vector of axiom preferences, along with its ordering in the original ranking. The axioms are operationalized as predicates that map a given document pair to a ternary ranking preference—prefer d_i, prefer d_j, or prefer neither. The explanation model aggregates a vector of such axiomatic ranking preferences to a final ordering for a given document pair in such a way that the ranking decisions of the retrieval model under investigation are reproduced as faithfully as possible. The explanation model's parameters give insights into how the retrieval axioms contribute to the ranking under scrutiny.

3.1 Operationalizing Retrieval Axioms

While a variety of axioms for different aspects of retrieval—such as ranking, evaluation, or relevance feedback—have been specified in the literature, we include only those that can be restated to express ranking preferences on pairs of documents. In operationalizing those axioms, we follow the approach of Hagen et al. [23], but make modifications to adapt their axiomatic re-ranking framework to our ranking explanation setting. In our setting, each axiom A implements a ternary predicate $A(d_i, d_j, q)$ that, given a document pair (d_i, d_j) and query q, maps to a ranking preference taking on values of 1, −1 or 0 depending on whether the axiom would rank document d_i higher, d_j higher, or has no preference on the pair.

Table 1 summarizes the retrieval axioms we employ in our study grouped by the general notion of relevance that they capture (axioms missing from the table cannot easily be restated to express actual ranking preferences). For example, the term frequency axioms TFC1, TFC3, and TDC constrain how the term frequency *tf* should manifest in document ranking and the first of these, TFC1, states that, given two documents of the same length and a single-term query, the document with more occurrences of the query term should be ranked higher [14].

The axioms are generally framed in artificial preconditions to allow for precisely reasoning about the properties of retrieval functions (e.g., TFC1 for documents with exactly the same length). However, since this limits their practical applicability, we make modifications following Hagen et al. [23]. For example, in case of TFC1, (1) we relax the equality constraint (i.e., we consider all document pairs with a length difference of at most 10%), (2) we strengthen the inequality constraint (i.e., requiring at least a 10% difference in term frequency), and (3) we generalize it to multi-term queries (i.e., using the sum of term frequencies over all query terms). The other term frequency axioms are modified in a similar way. As originally stated, given two equally discriminative query terms and two same-length documents, TFC3 prefers a document containing both terms over another that contains only one whereas for two query terms of differing discriminativeness, TDC prefers the document containing the more discriminative term [15]. While implementing term discriminativeness by inverse document frequency, we again relax the equalities and strengthen the inequalities as for TFC1 and generalize to summing over more query term pairs.

For the axioms capturing document length, lower-bounding, and query aspect constraints, we largely follow the operationalization of Hagen et al. [23]. In short, LNC1 prefers the shorter document given identical term frequency of all query terms, while TF-LNC prefers the document with more query term occurrences assuming the term frequencies of all non-query terms are the same. The lower-bounding axiom LB1 applies when there is a query term t such that both documents obtain the same retrieval score if t is removed from the query; in this case, LB1 prefers documents that contain t over those that do not [30]. The query aspect axiom REG evaluates the pairwise semantic similarity of the query terms, and prefers documents with more occurrences of the term that is least similar to all others—in our experiments, we employ the Wu-Palmer measure for term similarity [54]—whereas the axiom AND prefers documents that contain every query term at least once. The diversity axiom DIV prefers the document that is less similar to the query (measured via Jaccard similarity in our implementation).

The semantic similarity axiom STMC1 favors documents containing terms that are semantically more similar to the query terms, whereas STMC2 requires that a document exactly matching a query term once contributes to the score at least as much as matching semantically related terms arbitrarily many times instead [17]. We operationalize STMC1 and STMC2 based on the Wu-Palmer measure in the same way as Hagen et al. [23] but also additionally explore word embedding-based term similarity measures. We thus also have STMC1-f and STMC2-f that utilize 1 million fastText word vectors pre-trained with subword information on Wikipedia from 2017 [36] and STMC1-fr and STMC2-fr that utilize custom fastText embeddings trained on the Robust04 document collection. The term proximity axioms PROX1–PROX5 are employed in the same way as originally proposed by Hagen et al. [23]. Beyond these, Hagen et al. [23] also propose the axiom ORIG that simply reproduces the ranking preferences of the original retrieval model. While this axiom does not immediately apply to the explanation setting, we do include its retrieval model-specific versions in some experiments in order to study how one retrieval model's decisions might be explained in terms of those of another.

3.2 Aggregating Axiom Preferences

Once the axiomatic ranking preferences have been computed for a set of axioms and document pairs, we fit an explanation model to reconstruct the original ranking based on the axiomatic preferences. On the level of pairwise ranking decisions, this is a binary classification task. While a wide range of different models can accomplish

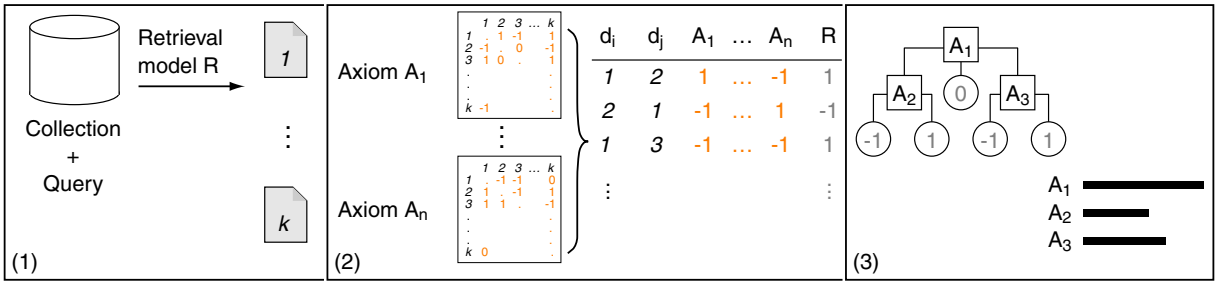

Figure 1: Overview of the axiomatic explanation pipeline. (1) The retrieval model R ranks a document set of size *k*. (2) Axioms produce ranking preferences for all document pairs. (3) A simple explanation model, trained to recreate the ordering produced by R using the axiom preferences as features, reveals which axioms generate the ranking.

this task, there is a clear case for using models that are simple enough to be inspected for insights on how exactly the retrieval axioms interact to generate the ranking under scrutiny. There is a trade-off between the fidelity of the explanation model—i.e., how well it reconstructs input rankings—and the degree to which it can be inspected: higher-capacity models tend to be less interpretable.

The goal of our present study is to test the feasibility of the axiomatic explanation approach and to examine the completeness of the axiom set. We thus employ a random forest model for axiom preference aggregation. This is at the higher-fidelity end of the spectrum but still offers useful, though limited, facilities for inspection in terms of feature importance. Once a more comprehensive axiom set is established—the need for this is indicated by our current results—, lower-capacity models with richer interpretability become reasonable. For instance, in a linear model like logistic regression, both the magnitude and signs of the parameter vector would be meaningful, while a single shallow decision tree could be interpreted as a Boolean formula that explains a given ranking.

4 EXPERIMENTAL SETUP

The experimental study described in this section focuses on two research questions: (1) To what extent can the axioms currently known from axiomatic IR faithfully reconstruct the decisions made by neural ranking models, and how does this compare to classical retrieval scoring functions? (2) Which retrieval axioms are most important in what scenarios, and what is the relationship between axiomatic explanations and ranking quality?

In order to answer these questions, we apply our axiomatic explanation framework on two standard evaluation datasets, and examine the top 1000 retrieved results across 200 different queries with three classical and five neural machine-learned retrieval models. For each ranking, we sample document pairs from the set of all pairwise orderings, and generate axiomatic ranking preferences. On these, we fit a random forest-based explanation model.

4.1 Collections

Neural ranking models need large amounts of labeled training data to produce rankings substantially superior to classical retrieval models [56], but not many training collections are available. We follow recent trends and use the Robust04 and MS MARCO datasets. Both are standard TREC collections containing relevance judgments and due to their distinctive size and characteristics—in terms of constituent documents and the associated search queries—allow us to test our approach in various scenarios.

The Robust04 dataset was developed to improve the consistency of retrieval techniques on difficult queries [52]. It offers a traditional TREC-style evaluation setup of a document collection (528,155 documents from the Financial Times, the Federal Register 94, the LA Times, and FBIS), a set of topics (250 topics with title and description representing information needs and a narrative detailing which documents are considered relevant), and manual relevance judgments from human assessors. We use the short keyword-style titles (e.g., “most dangerous vehicles”) as queries in our experiments and randomly sample 150 topics for training the neural rankers and the remaining 100 for our explainability experiments.

The MS MARCO dataset—originally only a collection for passage ranking and question answering—was recently used in the TREC 2019 and 2020 Deep Learning tracks since it provides large amounts of labeled data [38]. We use the data from the document retrieval task that provides an end-to-end retrieval scenario for full documents similar to Robust04. The dataset consists of 3.2 million web documents and 367.000 training queries. Relevance labels are derived from the labels of the associated passage retrieval dataset under the assumption that a document with a relevant passage is a relevant document. While being a potential source of label noise, this does not constitute a hindrance to our experimental study, since we are not interested in evaluating the effectiveness of the trained models, but in explaining how they reach their ranking decisions. We index the concatenation of the URL, the title, and the body text of each document and randomly sample 100 of the 200 official test queries from the document retrieval task of the Deep Learning track for our explainability experiments—training of the neural ranking models described in the next section. Note that the MS MARCO queries are notably longer (e.g., “what are the effects of having low blood sugar”) than the Robust04 queries.

4.2 Ranking Models

In our study, we apply those neural ranking models on each dataset that have been prominently applied in the respective setting in previous work [11, 21, 34, 42, 57]—leaving a larger study with all rankers on both datasets for future work. For comparison, we also add three classical retrieval models.

On Robust04, we study three pairwise neural ranking models. (1) MatchPyramid [42] uses a query-document interaction matrix as input to a convolutional neural network to extract matching patterns; we employ the cosine similarity variant (referred to as MP-COS henceforth), which performed best in preliminary experiments. (2) DRMM [21] deploys a feed-forward neural network over query-document count histograms to output a relevance score. (3) PACRR-DRMM [34] combines the PACRR [24] model (which uses CNNs with different kernel sizes to extract position-aware signals from n-gram patterns) with the aggregation of DRMM.

The neural models were set to re-rank the top-1000 results of BM25. We trained DRMM and MP-COS using the MatchZoo toolkit [22], fine-tuning according to the hyperparameters reported in the respective publications, and for PACRR-DRMM, we used its authors' implementation.[2] For all models, we used word embeddings fitted on Robust04, and labeled training data from the TREC Robust track. Our explainability experiments are run on the result sets of the 100 queries held out from the ranking model training.

On MS MARCO, we study two BERT-based ranking models. (1) DAI [11] is trained by fine tuning BERT to predict passage level relevance; we use the MaxP variant, where document relevance is the maximum of its passage-level relevances (DAI-MAXP). (2) BERT-3S [57] aggregates top-k sentence scores, evaluated by a transfer model,[3] to compute document relevance scores. Both models fine-tune a pretrained BERT model during training, and then use an aggregation procedure to predict document relevance; for more details on their training, we refer readers to the respective original publications. For our experiments, we used the MS MARCO training set from the TREC 2019 Deep Learning track, and chose the best performing models according to MAP calculated on the corresponding validation set. Our explainability experiments are run on the results sets of 100 queries from the MS MARCO test set.

For the explainability experiments, we complement the neural ranking models with the classical retrieval models BM25, TF-IDF, and PL2—all parameters set to their defaults according to the implementation in the Anserini toolkit [55].

4.3 Explanation Parameters

For our experiments, we instantiate the axiomatic explanation framework described in Section 3 with 20 axioms: the 16 axioms shown in Table 1 and the aforementioned 4 embedding-based variants for the semantic-similarity axioms STMC1 and STMC2. We operate on the top-1000 results of each ranking, but sample only a subset of all constituent document pairs. Based on the assumption that explanations of the top ranks are of the most interest, we follow a non-uniform sampling strategy that includes all pairs of documents from the top-20 ranks, plus 1% of all remaining pairs sampled uniformly at random. We instantiate the explanation model as a random forest with 128 trees of maximum depth 20.

The input to the explanation model is a set of pairwise ranking decisions made by the retrieval model to be explained, at a scope that depends on the granularity of the explanation model (Section 4.4). This input dataset is randomly split into ten folds which are alternately used to fit the explanation model, and to evaluate its explanation fidelity, in a standard ten-fold cross-validation setting. Every individual instance comprises the identifiers of the two documents involved, the ranking preferences for this pair for each of the 20 axioms, and the ranking preference of the retrieval model to be explained. All experiments use the latter ranking preference as the dependent variable for the explanation model to predict and the axiomatic ranking preferences as independent variables. Note that if an instance for document pair (d_i, d_j) is included, so is the instance for pair (d_j, d_i)—with inverted preferences. Both instances forming such mirrored pairs are always assigned to the same cross-validation fold to avoid train-test information leakage.

4.4 Explanation Model Locality

To answer our research questions related to the degree of locality at which axiomatic explanations apply, we train explanation models not only at the scope of the full ranking, but also at the scope of subsets of the ranking. Overall, we consider three types of locality: (1) locality by query, (2) locality by ranking position, and (3) locality by score differences. The distinction between ranking and score difference is useful to incorporate a notion of degree of certainty of the ranking model—a ranking assigns documents to different positions even if they obtain the same score. We create 24 bins of approximately the same size for locality by ranking position and for locality by score differences, and we combine the these with locality by query to obtain five binning strategies in total that we use to select documents to train our explanation models.

The left-hand side of Table 2 illustrates five different experiment configurations resulting from this setup, which fit explanation models at the following scopes: (1) one model per query, yielding 100 explanation models per retrieval model and dataset; (2) one model per one of the 24 bins of the ranking differences across all queries; (3) one model per one of the 24 bins of min-max normalized differences in retrieval score across all queries; (4) one model per combination of query and rank-difference bin; (5) one model per combination of query and score-difference bin. The first two columns of Table 2 ("Scope" and "Per Retrieval Model") show the scopes of the explanation models as described above, along with the number of explanation models per retrieval model resulting from the respective granularity level. The next two columns ("Train" and "Test") show the average number of training and test instances per cross-validation fold for each of these explanation models. Note that differences—mostly in the number of the most finely-granular explanation models—arise from the fact that not every retrieval model returns the full top-1000 results for every query. The values shown in the table are averaged over all retrieval models.

5 RESULTS

The "Explanation Fidelity" columns of Table 2 show to what extent our axiomatic explanation framework can characterize three classical retrieval models (BM25, TF-IDF, and PL2) and several neural rankers on the Robust04 and MS MARCO datasets. Explanation fidelity is measured as the accuracy of the explanation model in terms of reproducing the retrieval model's ranking decisions, macro-averaged over the ten cross-validation folds and over the number of explanation models (column "per Retrieval Model" in Table 2).

[2] https://github.com/nlpaueb/deep-relevance-ranking
[3] We fine-tuned this model using MS MARCO; Microblogs were not used.

Table 2: Overview of the ranking explanation experiments. Explanation fidelity is measured as the proportion of document pairs ordered correctly, and is averaged over ten cross-validation folds across all models evaluated in the corresponding row.

Explanation Models		Instances per Model		Explanation Fidelity					
Scope	per Retr. Model	Train	Test	Classical Retrieval Models			Neural Retrieval Models		
Robust04				**BM25**	**TF-IDF**	**PL2**	**MP-COS**	**DRMM**	**PACRR-DRMM**
query	100	8,943	1,279	**0.75**	**0.66**	0.78	**0.67**	**0.68**	**0.72**
rank-diff bin	24	38,213	4,380	0.71	0.63	0.77	0.59	0.61	0.67
score-diff bin	24	38,327	4,265	0.72	0.64	0.78	0.59	0.61	0.68
query, rank-diff bin	2,368	384	44	0.73	0.64	0.77	0.65	0.66	0.70
query, score-diff bin	2,394	383	44	0.74	0.65	**0.79**	0.64	0.66	0.70
MS MARCO				**BM25**	**TF-IDF**	**PL2**		**BERT-3S**	**DAI-MAXP**
query	100	8,936	1,278	**0.64**	**0.60**	**0.63**		**0.61**	**0.59**
rank-diff bin	24	38,208	4,350	0.60	0.56	0.59		0.57	0.54
score-diff bin	24	38,280	4,278	0.61	0.56	0.59		0.59	0.55
query, rank-diff bin	2,400	382	44	0.62	0.58	0.61		0.60	0.57
query, score-diff bin	2,376	386	44	0.63	0.60	0.62		0.61	0.58

5.1 Overall Explanation Fidelity

We achieve explanation fidelities of 0.54 and higher for all examined ranking models under all considered parameters. The classical retrieval model PL2 achieves the best explainability of nearly 80% on the Robust04 dataset, where the explanation granularity makes little difference. This high accuracy indicates that the limited set of simple retrieval axioms—although individually comprehensible for humans—can be combined to explain at least some of these relatively complex ranking formulas. The BM25 model is nearly as accurately explainable as PL2, even though BM25 can be considered more complicated since it has two tunable parameters, and PL2 has none. Rankings produced with TF-IDF illustrate that the simplicity of the ranking model does not guarantee good explainability. The accuracy of the TF-IDF explanations varies only slightly over different explanation model scopes, which is an observation that repeats for all classical ranking models under consideration.

By contrast, the accuracy often varies more for different explanation scopes in case of rankings produced by neural models. We find that this does not pose a problem, since the simple query-level granularity is always a reasonable choice that often outperforms more complicated setups like binning by rank difference. In the end, the scope of the explanation model does not strongly influence the fidelity of the explanations for any retrieval model. Fitting one explanation model per ranking appears to be the most straightforward approach, and in most cases the best-performing one.

The explanation fidelity of all neural ranking models under investigation on Robust04 is within the range of the fidelities obtained for the classical models TF-IDF and BM25. PACRR-DRMM reaches an accuracy of 0.72, almost at the level of BM25. Similarly, rankings from MP-COS and DRMM obtain accuracies slightly above TF-IDF, but both almost double the accuracy difference across model scopes.

On MS MARCO, the explainability for all classical retrieval models is much lower than on Robust04, especially for PL2. As outlined in Section 4.1, the two collections significantly vary in terms of size and query characteristics, which may explain the drop in fidelity. The neural rankers tested on MS MARCO also achieve poorer explanation fidelities compared to those tested on Robust04, but are not directly comparable. The BERT-3S model attains slightly better explanations than DAI-MAXP across all explanation model scopes. For all retrieval models tested on MS MARCO, the query-scope explanation models perform best.

Due to our sampling strategy, the explanation model solves a balanced binary classification problem, where an accuracy of 0.5 corresponds to failure to explain the given ranking decisions better than random chance. As a sanity check, we apply the explanation model also to randomly-shuffled variants of the rankings from the previous experiment. In this setting, the explanation fidelity remains consistently below 0.52 everywhere. Thus, our axiomatic framework explains all retrieval models at least somewhat better than random chance. However, especially for the MS MARCO rankings, explanation fidelity is very limited. Going forward, we investigate the aggregated results from Table 2 in finer detail to better understand in what contexts the axiomatic explanation models perform well, and which axioms are responsible.

5.2 Explanation Fidelity by Score Difference

An initial comparison of the explanation models at rank-difference scope to those at score-difference scope indicates that the score difference is slightly more useful both when training explanation models as well as when applying those models to explain rankings. This is expected, as a difference in rank does not necessarily correspond to a high confidence of the ranking model in the difference, since it does not take score ties into account [27]. For this reason, we proceed with the score-difference based binning.

To better understand how locality within the ranking might affect the fidelity of our axiomatic explanations, we inspect the performance of the query-level explanation models on the level of individual bins of the pairwise score differences produced by the studied retrieval models. For each dataset and retrieval model, we employ the 100 explanation models of the "query" scope (cf. Table 2) and subdivide the document pair instances from the test folds into 24 bins according to the min-max normalized difference in

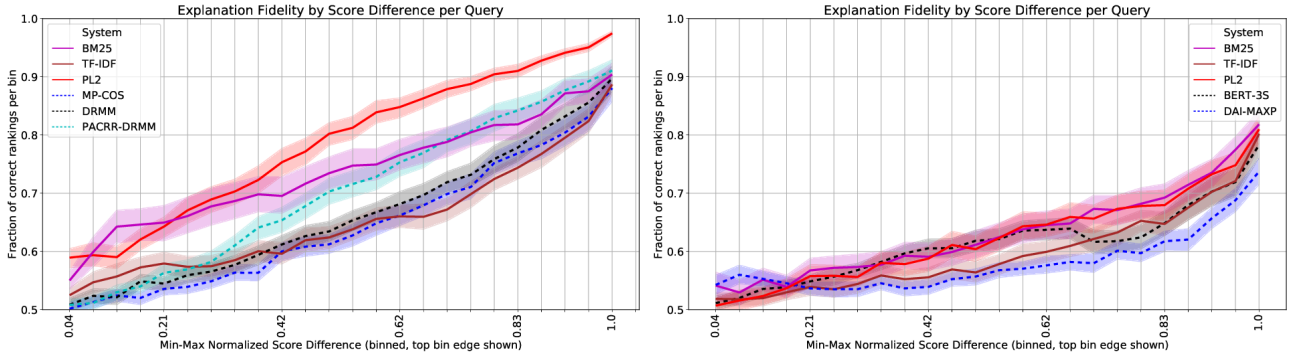

Figure 2: Explanation fidelity by score difference per query on Robust04 (left) and MS MARCO (right).

retrieval score. The bin edges are chosen in such a way that all bins contain the same number of instances on average across queries. We then compute the explanation fidelity for each bin separately and macro-average the results across the testing folds and queries.

Figure 2 shows the explanation fidelity by the score difference binning per query as one line for each retrieval model's averaged explanation fidelities. Solid lines represent the classical retrieval models, and dashed lines the neural ones. The bands indicate 95% confidence intervals across queries. Small changes in a retrieval model's scores are hardly ever well explainable. This indicates that if the retrieval model is uncertain about a document pair and assigns a small score difference, this pair will also be hard to explain; large score differences, by contrast, will be much more explainable.

The increase in explainability with increasing score difference is more pronounced for Robust04; the explanations on both datasets perform pretty poorly at the low score difference end. This indicates that the better explainability of Robust04 rankings (Table 2) mainly results from document pairs far apart in retrieval score, while for the larger MS MARCO collection the probably much larger clusters of highly-similarly scored documents may be more problematic.

For a deeper look into what the axiomatic preferences actually explain, we examine the feature importance of the axioms in the random forest explanation model. To this end, we fit the explanation models on four more coarse-grained bins over the min-max normalized retrieval scores, again chosen to contain approximately the same number of instances. In this setting, we examine which axioms account for small score differences of 12.6% or less, middling differences from 12.6% to 25.2%, and from 25.6% to 50%, as well as large differences greater than 50%. We perform this analysis on the Robust04 dataset due to its better overall explainability. To quantify the feature importance of an axiom A, we measure the mean decrease in Gini impurity averaged over all decision trees in the ensemble, and over all splits where A is used, weighted by the number of training instances involved in the split [29].

Table 3 shows the mean decrease in impurity of the three most important axioms for each combination of score difference bin and retrieval model. Most strikingly, we found a large degree of overlap among nearly all investigated retrieval models, and across the majority of the score difference bins. The query aspect axiom REG features prominently; while this axiom rewards occurrence of the most "outlier" query term in result documents, the mere fact that it directly rewards query term occurrence may suffice to make it feature prominently in axiomatic explanations. The term proximity axiom PROX4 contributes to the axiomatic explanations in almost all cases. As originally formulated [23], PROX4 rewards the occurrence of close groupings of all query terms with few non-query terms in between. None of the classical retrieval models incorporate notions of term proximity, but meeting the PROX4 constraint implies that a document contains all query terms. The diversity axiom DIV, which penalizes Jaccard similarity to the query, constitutes a useful feature for most retrieval models and score differences, but since the explanation models are based on decision trees, they will likely use violations of the DIV axiom as a positive signal, as this is more in line with the behavior of the classical retrieval models. By further simplifying the explanation model to, for example, a regression model, such effects will be apparent in the algebraic signs of the parameters.

The rankings produced by the PL2 retrieval model, which tend to be best explainable overall on Robust04, tend to match the same axioms as other retrieval models for small score differences. However, for large score differences of 50% and more, PL2 has no overlap among the top three axioms with any other model, while all other models share the same top three axioms in this range. Interestingly, two variants of the semantic matching axiom STMC1 feature prominently for PL2 in this range, even though the PL2 scoring function considers exact matches only. It seems that the document property rewarded by STMC1—the presence of terms semantically similar to query terms—correlates with the divergence from randomness of the term frequency of query terms, which PL2 measures.

5.3 Ranking Quality and Explanation Fidelity

Over all Robust04 queries, we observe weak but consistently positive correlations between retrieval model effectiveness as measured by nDCG and the explanation fidelity of the query-scoped explanation models (Spearman's correlations of 0.16 for DRMM, 0.26 for MP-COS, and 0.2 for PACRR-DRMM). Table 4 shows some examples: the good rankings for queries like 425 or 612 tend to be much better explainable than the worse rankings for queries like 344 that are not well-explainable. However, there are notable exceptions like query 367 on which no system performs particularly well in terms of nDCG but the explanation fidelity is high across all models.

Table 3: Axiom importance for explanation fidelity on ROBUST04 in terms of mean decrease in impurity for the top three axioms across min-max normalized score difference bins (PACRR-DRMM denoted as PACRR for brevity).

	Diff. ≤ 12.6%					12.6% < **Diff.** ≤ 25.2%				
	REG	PROX4	LNC1	DIV	LB1	REG	PROX4	LNC1	DIV	
BM25	0.14	0.13	0.17			0.16	0.16	0.13		
TF-IDF	0.15	0.12			0.12	0.15	0.12	0.15		
PL2	0.15	0.13		0.11		0.14		0.15	0.13	
MP-COS	0.16	0.11		0.11		0.17	0.11		0.11	
DRMM	0.16	0.11		0.11		0.16	0.11		0.11	
PACRR	0.14	0.12		0.12		0.15	0.14		0.11	

	25.2% < **Diff.** ≤ 50%					50% < **Diff.** ≤ 100%					
	REG	PROX4	PROX5	LNC1	DIV	REG	PROX4	PROX5	DIV	STMC1-f	STMC1-fr
BM25	0.21	0.16	0.09			0.28	0.16	0.15			
TF-IDF	0.16	0.13		0.11		0.23	0.14	0.15			
PL2	0.11			0.11	0.20				0.25	0.13	0.13
MP-COS	0.20	0.13			0.11	0.31	0.14	0.11			
DRMM	0.19	0.12	0.11			0.28	0.14	0.18			
PACRR	0.20	0.15	0.11			0.29	0.15	0.17			

The top half of Table 5 shows the first ten results of the DRMM ranking for query 367, along with our axiomatic explanations illustrated as a summary aggregating how many axioms would rank a particular result document in the same position, higher, or lower compared to the system ranking. Note that no axiom expresses a preference for every document pair—in fact, the total number of preferences is usually way lower than the actual number of 20 axioms. The last column illustrates the ranking preferences of individual axioms, as they relate to explaining the position of the document in the current row. Out of all axioms with any ranking preference regarding the row's document, we show the top three by how much they support or oppose the document's rank.

The many non-relevant top-10 results for query 367 can be explained by the specifics of the TREC topic. While DRMM only saw the single-term query "piracy," the assessor instructions clarify that traditional high-seas piracy is meant, but not computer piracy as covered in many of the top-10 documents. The DRMM ranking is best explained by the REG and DIV axioms that may not really be the most "intuitive" choices for single-term queries. Still, REG degenerates to simply preferring documents with more query term occurrences for single-term queries which then "makes sense" for query 367. However, the agreement between the axioms and the top-10 ranks is rather poor; the overall good explanation fidelity comes from more distant pairs beyond the top ten results.

The lower half of Table 5 shows the top-10 DRMM results for the two-term query 425 "counterfeiting money". Here, DRMM performs considerably better and there also is a better agreement between the axioms and the top-10 ranks—with the term proximity axioms being the most important to explain the ranker's decisions.

5.4 Limitations and Summary

The results of the preceding analysis indicate that while good explanation fidelity is possible in some contexts, the reason why axiomatic explanations work at all can be somewhat incidental to the criteria employed by the retrieval model to be explained. In this sense, not much can yet be said on what brings about the ranking decisions of the more inscrutable neural rankers, except that similar criteria to classical retrieval scenarios clearly apply. However, we also do find evidence that something else is at play. In an experiment with versions of the ORIG axiom [23] expressing the ranking preferences of the classical retrieval models, all the classical models are highly effective at explaining each other's decisions. The full axiom set of the 20 before mentioned axioms extended by two axioms each expressing the rank preferences of the respective two not-to-be-explained classical models reaches explanation fidelities above 99% for the classical models. Still, the explanation fidelity for the neural models increases only moderately.

Table 4: Explanation fidelity and retrieval effectiveness for selected Robust04 queries (PACRR-DRMM as PACRR).

Query	**Explanation fidelity**			**Topic title**	**nDCG**		
	MP-COS	DRMM	PACRR		MP-COS	DRMM	PACRR
344	0.55	0.57	0.60	Abuses of E-Mail	0.27	0.26	0.33
352	0.56	0.64	0.61	British Chunnel impact	0.17	0.15	0.15
356	0.73	0.61	0.60	Postmenopaus. estrogen Britain	0.10	0.15	0.10
367	0.74	0.82	0.87	Piracy	0.34	0.29	0.31
399	0.61	0.57	0.71	Oceanographic vessels	0.28	0.30	0.28
409	0.84	0.81	0.62	Legal, Pan Am, 103	0.24	0.35	0.44
425	0.78	0.81	0.85	Counterfeiting money	0.74	0.77	0.75
612	0.80	0.81	0.85	Tibet protesters	0.46	0.57	0.54
618	0.81	0.82	0.83	Ayatollah Khomeini death	0.52	0.41	0.44
684	0.55	0.67	0.66	Part-time benefits	0.26	0.41	0.39

We thus hypothesize that formulations we did use for the 20 axioms still lack some reliable indicators for some basic relevance signals used by classical retrieval models. For instance, even though there are axioms capturing term frequency, their current formulation might not be very useful in practice. To test this, we investigate how often the individual axioms' preconditions are satisfied, and find that the 10% relaxation of equality constraints—we used it following Hagen et al. [23]—may still be too strict. In an experiment on Robust04, we find that the term frequency axioms' document length precondition is satisfied only in approximately 7% of the document pairs, the preconditions of LNC1 only in 9%, the proximity axioms PROX1–3 can be applied to only 21%, and axiom LB1 to only 36% of the document pairs, while the remaining axioms apply to 90% or more of the document pairs.

In summary, we find that large differences in retrieval score can be reasonably well explained with the simple axiomatic feature set employed in our study. Especially for the MS MARCO dataset, the explainability does not depend very much on the specific retrieval model used to produce the ranking, and across all datasets, the simple setup of training one explanation model per query outperforms more complicated binning approaches, although binning may still be useful to understand model behavior across levels of score differences. However, a closer investigation of this behavior indicates that our current axiom set does not fully capture the scoring criteria of most ranking models, one possibly reason being the strict preconditions contained in several of the axioms.

6 CONCLUSION & FUTURE WORK

We have introduced an axiomatic framework to explain the result rankings of information retrieval systems in terms of how well a system's ranking decisions adhere to a set of axiomatic constraints.

Table 5: Example rankings with axiomatic explanations. (Non-)relevant Document IDs have a suffix + (−). Axiomatic explanations show how many axioms would rank this result the same (⇔), higher (↑) or lower (↓), followed by up to three most relevant axioms, and how many other results they rank the same (superscript) or differently (subscript).

System: DRMM	**Query:** 367 "piracy"	**Axiomatic explanation**	
Rank	**Docid**	**Content**	⇔ / ↑ / ↓	**Most relevant axioms**
1	FT944-16684$^-$	Software companies offer rewards in anti-piracy drive – Leading software compan…	1 / 0 / 1	REG^{9}_{0} DIV^{3}_{6}
2	FT931-8281$^-$	Survey of Personal and Portable Computers (21): Tougher times for pirates / a lo…	2 / 0 / 1	REG^{8}_{1} DIV^{4}_{5} $LNC1^{1}_{0}$
3	FT934-14966$^-$	Sixties buccaneer deals in high-tech piracy – In his days as a radio caroline p…	3 / 0 / 3	DIV^{4}_{5} REG^{3}_{5} $STMC2^{1}_{0}$
4	FT911-1567$^-$	A rich haul from the sound of music: the illicit copying and sale of recorded mu…	3 / 1 / 0	REG^{7}_{2} DIV^{4}_{5} $LNC1^{1}_{0}$
5	LA032889-0045$^-$	Busting cable pirates; simi valley piracy case is one of the first to result in …	3 / 1 / 0	REG^{6}_{1} DIV^{4}_{5} $LNC1^{2}_{0}$
6	FBIS3-43017$^-$	Computer piracy in russia is a widespread phenomenon. the world average ratio of…	5 / 1 / 0	REG^{6}_{1} DIV^{3}_{5} $LNC1^{1}_{0}$
7	FBIS3-42979$^-$	Computer piracy in russia is a widespread phenomenon. the world average ratio of…	2 / 1 / 0	REG^{6}_{1} DIV^{3}_{5} $LNC1^{1}_{0}$
8	FT923-9880$^+$	Jakarta sinks plan to combat piracy – Plans for an international centre to figh…	5 / 0 / 0	REG^{8}_{1} DIV^{8}_{1} $STMC2^{2}_{0}$
9	FT944-9277$^-$	UK company news: BSkyB says piracy could undermine float confidence – British s…	2 / 1 / 0	DIV^{8}_{1} REG^{8}_{0} $LNC1^{0}_{1}$
10	FT924-15875$^-$	Piracy warnings – A 24-hour centre to counter piracy in the seas of south-east …	2 / 0 / 0	DIV^{9}_{0} REG^{9}_{0}

System: DRMM	**Query:** 425 "counterfeiting money"		**Most relevant axioms**
1	FBIS3-58171$^+$	The head of the gang that wanted to circulate some 970,000 counterfeit dollars i…	6 / 0 / 5	$PROX5^{8}_{1}$ $PROX3^{7}_{2}$ $STMC1\_f^{7}_{2}$
2	FBIS4-46741$^+$	Crime Counterfeiting is probably one of the world's oldest and most widespread t…	10 / 0 / 1	$PROX2^{7}_{2}$ $PROX3^{7}_{2}$ REG^{7}_{2}
3	FBIS4-26260$^+$	The Hongqiao District people's court examined and concluded a case of traffickin…	5 / 0 / 5	$PROX4^{7}_{2}$ $PROX5^{7}_{2}$ $PROX2^{7}_{2}$
4	LA091590-0091$^+$	PLUMBERS DISCOVER CASH FLOW PROBLEM IN SEWER; COUNTERFEITING:…	7 / 2 / 0	$PROX4^{7}_{2}$ $PROX5^{7}_{2}$ $PROX1^{6}_{3}$
5	FBIS3-54773$^+$	On 25 December a criminal gang of six Chechnya inhabitants was arrested in St. P…	7 / 1 / 2	$PROX4^{7}_{2}$ $PROX5^{7}_{2}$ REG^{5}_{4}
6	LA102189-0077$^+$	CALIFORNIA IN BRIEF; MODESTO; MAN INDICTED IN COUNTERFEITING…	7 / 3 / 0	$PROX4^{7}_{1}$ $PROX5^{6}_{3}$ $PROX1^{6}_{3}$
7	FBIS4-47199$^+$	The number of counterfeit ruble bank notes, bank notes of convertible currency, …	7 / 4 / 0	$PROX5^{8}_{1}$ $STMC1^{6}_{3}$ $PROX4^{6}_{1}$
8	FBIS4-58263$^+$	Counterfeit, 1990-issue $100 bills have recently found their way onto the Jorda…	6 / 4 / 0	$PROX4^{7}_{2}$ REG^{7}_{2} $PROX3^{7}_{1}$
9	FBIS4-59139$^+$	For some time, Tel Aviv has anxiously been raising with Egyptian politicians an …	10 / 1 / 0	$PROX1^{8}_{1}$ $PROX3^{8}_{1}$ $PROX2^{8}_{1}$
10	LA010390-0055$^-$	RIVAL ATHLETIC SHOE MAKERS FORMED AN ALLIANCE TO BATTLE COUNT…	6 / 5 / 0	$PROX1^{9}_{0}$ $PROX4^{8}_{1}$ $PROX5^{8}_{1}$

Instantiated with a set of 20 axioms from the literature and a random forest model to reconstruct pairwise orderings from axiomatic ranking preferences, we have demonstrated our suggested framework's general capacity to explain rankings in an experimental study on the Robust04 and MS MARCO test collections. The results show that axiomatic explanations for eight different retrieval systems—five of them complex deep neural network-based ranking functions and three classical scoring functions—work reliably for document pairs with very different retrieval scores (i.e., corresponding to a high confidence in a difference in relevance). Pairs with more similar retrieval scores are more difficult to explain—not too surprising given the rather few retrieval aspects that the 20 axioms do cover. Especially axioms with a precondition constraining the documents' length difference can rarely be applied, even when this constraint is relaxed to allow for a 10% difference, as was suggested in previous studies. Further relaxing or even dropping preconditions entirely may be an easy remedy, but also a vast departure from the original axioms and their formalization of the constraints they capture. Instead, it seems desirable to formulate new axioms that capture the same ideas in a more practically applicable way, or that capture retrieval constraints not yet covered by the known axioms, and to develop a weighting scheme that can quantify the degree to which preconditions are satisfied.

The explanation fidelity on the smaller, more genre-focused Robust04 collection with its shorter queries is superior to that on the MS MARCO dataset. Further investigation into the causes of this discrepancy is warranted, but the vastly different characteristics of the respective queries and documents seem likely candidates. The explainability of neural rankers is mostly on par with that of classical retrieval functions, and there are notable overlaps in the axioms that are most useful to the explanation models. Still, the "known" and studied axioms do not cover a range of aspects important to modern search engines such as the timeliness of results, stylistic and readability considerations, or how well some results match user preferences expressed through previous interactions.

While formalizing and operationalizing axiomatic constraints for such properties certainly seems worthwhile, such an endeavor was beyond the scope of our paper. Even though we can demonstrate promising first steps to axiomatically explain retrieval systems' result rankings, the addition of further well-grounded axiomatic constraints capturing other retrieval aspects seems to be needed to further improve the explanations. Its current limitations notwithstanding, we consider our approach a promising complement to the more tightly-controlled studies from previous work [7, 32, 44]. While the latter shed light on the general principles under which complex relevance scoring models operate, our axiomatic reconstruction framework could help IR system designers—or even end users—make sense of a concrete ranking for a real-world query.

ACKNOWLEDGMENTS

This work has been partially supported by the DFG through the project "ACQuA: Answering Comparative Questions with Arguments" (grant HA 5851/2-1) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999). Jaspreet Singh's contributions were made prior to his affiliation with Amazon.

REFERENCES

[1] Alon Altman and Moshe Tennenholtz. 2005. Ranking Systems: The PageRank Axioms. In *Proceedings of EC 2005*. 1–8.

[2] Enrique Amigó, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. 2017. Axiomatic Thinking for Information Retrieval: And Related Tasks. In *Proceedings of SIGIR 2017*. 1419–1420.

[3] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. 2013. A General Evaluation Measure for Document Organization Tasks. In *Proceedings of SIGIR 2013*. 643–652.

[4] Siddhant Arora and Andrew Yates. 2019. Investigating Retrieval Method Selection with Axiomatic Features. In *Proceedings of the AMIR Workshop at ECIR 2019*. 18–31.

[5] Peter Bruza and Theo W. C. Huibers. 1994. Investigating Aboutness Axioms using Information Fields. In *Proceedings of SIGIR 1994*. 112–121.

[6] Luca Busin and Stefano Mizzaro. 2013. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *Proceedings of ICTIR 2013*. 8.

[7] Arthur Câmara and Claudia Hauff. 2020. Diagnosing BERT with Retrieval Heuristics. In *Proceedings of ECIR 2020*. 605–618.

[8] Ronan Cummins and Colm O'Riordan. 2007. An Axiomatic Comparison of Learned Term-Weighting Schemes in Information Retrieval: Clarifications and Extensions. *Artif. Intell. Rev.* 28, 1 (2007), 51–68.

[9] Ronan Cummins and Colm O'Riordan. 2011. Analysing Ranking Functions in Information Retrieval Using Constraints. In *Information Extraction from the Internet*, 241–256.

[10] Ronan Cummins and Colm O'Riordan. 2012. A Constraint to Automatically Regulate Document-Length Normalisation. In *Proceedings of CIKM 2012*. 2443–2446.

[11] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of SIGIR 2019*. 985–988.

[12] Fan Ding and Bin Wang. 2008. An Axiomatic Approach to Exploit Term Dependencies in Language Model. In *Proceedings of AIRS 2008*. 586–591.

[13] Hui Fang. 2008. A Re-Examination of Query Expansion Using Lexical Resources. In *Proceedings of ACL 2008*. 139–147.

[14] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A Formal Study of Information Retrieval Heuristics. In *Proceedings of SIGIR 2004*. 49–56.

[15] Hui Fang, Tao Tao, and ChengXiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29, 2 (2011), 7:1–7:42.

[16] Hui Fang and ChengXiang Zhai. 2005. An Exploration of Axiomatic Approaches to Information Retrieval. In *Proceedings of SIGIR 2005*. 480–487.

[17] Hui Fang and ChengXiang Zhai. 2006. Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In *Proceedings of SIGIR 2006*. 115–122.

[18] Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. 2019. A Study on the Interpretability of Neural Retrieval Models using DeepSHAP. In *Proceedings of SIGIR 2019*. 1005–1008.

[19] Shima Gerani, ChengXiang Zhai, and Fabio Crestani. 2012. Score Transformation in Linear Combination for Multi-Criteria Relevance Ranking. In *Proceedings of ECIR 2012*. 256–267.

[20] Sreenivas Gollapudi and Aneesh Sharma. 2009. An Axiomatic Approach for Result Diversification. In *Proceedings of WWW 2009*. 381–390.

[21] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-Hoc Retrieval. In *Proceedings of CIKM 2016*. 55–64.

[22] Jiafeng Guo, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2019. MatchZoo: A Learning, Practicing, and Developing System for Neural Text Matching. In *Proceedings of SIGIR 2019*. 1297–1300.

[23] Matthias Hagen, Michael Völske, Steve Göring, and Benno Stein. 2016. Axiomatic Result Re-Ranking. In *Proceedings of CIKM 2016*. 721–730.

[24] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A Position-Aware Neural IR Model for Relevance Matching. In *Proceedings of EMNLP 2017*. 1049–1058.

[25] Maryam Karimzadehgan and ChengXiang Zhai. 2012. Axiomatic Analysis of Translation Language Model for Information Retrieval. In *Proceedings of ECIR 2012*. 268–280.

[26] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of KDD 2016*. 1675–1684.

[27] Jimmy Lin and Peilin Yang. 2019. The Impact of Score Ties on Repeatability in Document Ranking. In *Proceedings of SIGIR 2019*. 1125–1128.

[28] Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling Trends in Machine Learning Scholarship. *Queue* 17, 1 (2019), 45–77.

[29] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. 2013. Understanding Variable Importances in Forests of Randomized Trees. In *Proceedings of NIPS 2013*. 431–439.

[30] Yuanhua Lv and ChengXiang Zhai. 2011. Lower-Bounding Term Frequency Normalization. In *Proceedings of CIKM 2011*. 7–16.

[31] Yuanhua Lv and ChengXiang Zhai. 2012. A Log-Logistic Model-Based Interpretation of TF Normalization of BM25. In *Proceedings of ECIR 2012*. 244–255.

[32] Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2020. ABNIRML: Analyzing the Behavior of Neural IR Models. *arXiv:2011.00696 [cs]* (2020).

[33] Brian P. McCune, Richard M. Tong, Jeffrey S. Dean, and Daniel G. Shapiro. 1985. RUBRIC: A System for Rule-Based Information Retrieval. *IEEE Trans. Software Eng.* 11, 9 (1985), 939–945.

[34] Ryan McDonald, George Brokos, and Ion Androutsopoulos. 2018. Deep Relevance Ranking using Enhanced Document-Query Interactions. In *Proceedings of EMNLP 2018*. 1849–1860.

[35] Carlo Meghini, Fabrizio Sebastiani, Umberto Straccia, and Costantino Thanos. 1993. A Model of Information Retrieval Based on a Terminological Logic. In *Proceedings of SIGIR 1993*. 298–307.

[36] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of LREC 2018*.

[37] Bhaskar Mitra and Nick Craswell. 2017. Neural Text Embeddings for Information Retrieval. In *Proceedings of WSDM 2017*. 813–814.

[38] Bhaskar Mitra and Nick Craswell. 2019. Duet at TREC 2019 Deep Learning Track. In *Proceedings of TREC 2019*.

[39] Seung-Hoon Na, In-Su Kang, and Jong-Hyeok Lee. 2008. Improving Term Frequency Normalization for Multi-topical Documents and Application to Language Modeling Approaches. In *Proceedings of ECIR 2008*. 382–393.

[40] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Cognitive Computation Workshop at NIPS 2016*.

[41] Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage Re-Ranking with BERT. *arXiv:1901.04085 [cs]* (2020).

[42] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2016. A Study of MatchPyramid Models on Ad-hoc Retrieval. *arXiv:1606.04648* (2016).

[43] Razieh Rahimi, Azadeh Shakery, and Irwin King. 2014. Axiomatic Analysis of Cross-Language Information Retrieval. In *Proceedings of CIKM 2014*. 1875–1878.

[44] Daniël Rennings, Felipe Moraes, and Claudia Hauff. 2019. An Axiomatic Approach to Diagnosing Neural IR Models. In *Proceedings of ECIR 2019*. 489–503.

[45] Corby Rosset, Bhaskar Mitra, Chenyan Xiong, Nick Craswell, Xia Song, and Saurabh Tiwary. 2019. An Axiomatic Approach to Regularizing Neural Ranking Models. In *Proceedings of SIGIR 2019*. 981–984.

[46] Koustav Rudra and Avishek Anand. 2020. Distant Supervision in BERT-based Adhoc Document Retrieval. In *Proceedings of CIKM 2020*. 2197–2200.

[47] Jaspreet Singh and Avishek Anand. 2018. Posthoc Interpretability of Learning to Rank Models using Secondary Training Data. *arXiv:1806.11330* (2018).

[48] Jaspreet Singh and Avishek Anand. 2019. EXS: Explainable Search using Local Model Agnostic Interpretability. In *Proceedings of WSDM 2019*. 770–773.

[49] Jaspreet Singh and Avishek Anand. 2020. Model Agnostic Interpretability of Text Rankers via Intent Modelling. *Proceedings of FAT* 2020*. 618–628.

[50] Tao Tao and ChengXiang Zhai. 2007. An Exploration of Proximity Measures in Information Retrieval. In *Proceedings of SIGIR 2007*. 295–302.

[51] C. J. van Rijsbergen. 1986. A New Theoretical Framework for Information Retrieval. In *Proceedings of SIGIR 1986*. 194–200.

[52] Ellen M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *Proceedings of TREC 2004*.

[53] Hao Wu and Hui Fang. 2012. Relation Based Term Weighting Regularization. In *Proceedings of ECIR 2012*. 109–120.

[54] Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. In *Proceedings of ACL 1994*. 133–138.

[55] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In *Proceedings of SIGIR 2017*. 1253–1256.

[56] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In *Proceedings of SIGIR 2019*. 1129–1132.

[57] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. In *Proceedings of EMNLP-IJCNLP 2019*. 3488–3494.

[58] Dell Zhang, Robert Mao, Haitao Li, and Joanne Mao. 2011. How to Count Thumb-Ups and Thumb-Downs: User-Rating Based Ranking of Items from an Axiomatic Perspective. In *Proceedings of ICTIR 2011*. 238–249.

[59] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and Predict, and then Predict Again. In *Proceedings of WSDM 2021*. 418–426.

[60] Wei Zheng and Hui Fang. 2010. Query Aspect Based Term Weighting Regularization in Information Retrieval. In *Proceedings of ECIR 2010*. 344–356.