Alignment Rationale for Query-Document Relevance

Youngwoo Kim, Razieh Rahimi and James Allan
University of Massachusetts Amherst
Amherst, MA, USA
{youngwookim,rahimi,allan}@cs.umass.edu

ABSTRACT

Deep neural networks are widely used for text pair classification tasks such as as adhoc information retrieval. These deep neural networks are not inherently interpretable and require additional efforts to get rationale behind their decisions. Existing explanation models are not yet capable of inducing alignments between the query terms and the document terms – which part of the document rationales are responsible for which part of the query? In this paper, we study how the input perturbations can be used to infer or evaluate alignments between the query and document spans, which best explain the black-box ranker's relevance prediction. We use different perturbation strategies and accordingly propose a set of metrics to evaluate the faithfulness of alignment rationales to the model. Our experiments show that the defined metrics based on substitution-based perturbation are more successful in preferring higher-quality alignments, compared to the deletion-based metrics.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; Information retrieval query processing.

KEYWORDS

neural network explanation, document search, query highlighting, textual matching, text alignment, token-level explanation

ACM Reference Format:
Youngwoo Kim, Razieh Rahimi and James Allan. 2022. Alignment Rationale for Query-Document Relevance. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531883

1 INTRODUCTION

BERT-based neural network models have shown state-of-the-art performance in information retrieval tasks [2, 3, 11, 21]. However, due to their complex architectures, they have remained a black box and their underlying decision-making mechanisms are not clear, even to domain experts. There have been efforts to explain black-box models' behavior in terms of the input features (e.g., tokens in document ranking), either by assigning importance scores to the features or selecting a subset of features that are important to preserve the models decisions [5, 6, 10, 17]

Query:	Where is SIGIR 2022
Document:	SIGIR 2022 will be held in Madrid

Figure 1: An example alignment for the query span 'Where is'

However, we found few works that answer the alignment question:"If certain document tokens are important for relevance to the query, which part of the query do they respond to?" Figure 1 illustrates the goal of alignment. When exact match or soft match based ranking models were used, the alignment between query tokens and document tokens could be acquired with little additional effort. Such alignment information has also used to provide more information to users, such as summarizing and visualizing each of query terms appearances in long document [7, 8], also demonstrating the important of this alignment issue.

Acquiring alignment has two approaches: (1) aiming at building (ideally) 'correct' or useful alignments regardless of query-document scoring model, or (2) seeking an alignment that best explains (is faithful to) the model. We target the second approach here.

We investigate the possible uses of input perturbation approaches, which make no assumption about the model's internal architecture. If the model outputs different decisions for a perturbed instance (a small change to the inputs), we can expect that the changed features are somehow responsible for the model decisions. To expand feature importance to alignment explanation, one can test if importance of some document tokens depends on the existence of certain query terms. Unfortunately such complex perturbation is more likely to bring undesired consequences such as making the input text ungrammatical [6] or changing its meaning drastically such that the model's decision changes more than we would expect given small perturbations. For example, consider the case when the query is "Where is SIGIR 2022" and we want to test which parts of a document are responsible for each of "Where is" and "SIGIR 2022". If we remove "SIGIR 2022" from the query, the query becomes "Where is?". In the case of the BERT-based model trained on the MSMARCO dataset [12], the relevant documents for this reduced query are the ones that contain information about how the expression "Where is?" is used, instead of those that present a location of events or entities.

How often does that happen? Does that actually make the perturbation useless? Are there any fixes if it does? This study addresses these research questions.

The contributions of our paper are as followings:

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531883

(1) We propose perturbation-based metrics to evaluate alignment rationale for query-document relevance.[1]
(2) We investigate the behavior of the proposed metrics and demonstrate that they are mostly not strong enough to make binary decisions on alignment quality (good or bad), but they can be used to rank two alignment models.
(3) We propose that building perturbed instances that are more comparable to the instance being explained, is the key to improvement of evaluation metrics. We showed that a simple approach to get more comparable instances increases the metric coverage from 13% to 68%. [2]

2 RELATED WORKS

Jiang et al. [9] generate alignment rationale for the natural language inference task. The alignment is built and evaluated based on the attention mask. Specifically, the attention vector across two segments is removed if they are not in the alignment. There are two notable limitations. First, the method and evaluation is dependent on the specific architecture of the model. Second, in case of BERT-based models, the attention flow inside the same segment and the flow to special tokens ([CLS] or [SEP]) are always kept. Thus, the alignment could be built through these tokens even when the direct attention vectors are dropped.

Models for explaining information retrieval models can be categorized into two groups. First category relies on interpretable features such as exact match features [16, 17, 19]. The second category only selects (or assigns importance to) tokens of the documents as explanation and do not build explicit alignments between each of query terms and the selected document tokens [5, 14, 18, 20, 23]. Neither categories of the existing explanation models are directly applicable for building relevance alignments.

There are studies to understand the behavior of BERT- or transformer-based models by inspecting their attention weights [13, 22]. While the supposedly aligned tokens tend to have higher weights than the others, many of the weights might actually not change the model decision when removed [13]. Moreover, there are hundreds of different attention weights between any single token pairs, and how to combine them is yet an unsolved challenge.

In model explanation literature, sufficiency and necessity metrics are often used to measure faithfulness of explanations (rationales) [1, 4]. These two metrics do not penalize rationales for being too verbose and tend to favor longer rationales. When explanation models provide real-valued scores for input tokens, the rationales can be forced to be concise by selecting the top-k% of tokens as the final rationales [9], but this strategy is not applicable when the explanation models only provide binary decisions. In our work, we propose a modification of the *necessity* metric to control the verbosity, which is applicable even when only tokens are given binary scores.

3 ALIGNMENT RATIONALES

Let f be a black-box classifier model that given a query q and document d, returns the probability of d being relevant to q, i.e., $f(q,d) \rightarrow [0,1]$. We assume that the model predicts a document d as relevant to the query q if its output is higher than a pre-defined threshold θ_r, i.e., $f(q,d) \geq \theta_r$, and otherwise predicts it as non-relevant. We use $R(q,d) = 1$ to denote that document d is considered as relevant by the model f, and $R(q,d) = 0$ to denote the non-relevance prediction.

The prediction of model f for a given pair (q,d) can be explained in different formats depending on the desired goal for explanations. We focus on the explanation of text matching between the query and document as text matching has been shown to be a strong signal of relevance.

Assume that q and d are split into two sets of text spans $\mathcal{Q}$ and $\mathcal{D}$, respectively. The segmentation unit can be tokens, phrases, or sentences, and can vary for the query and document. We consider that each text span of the query indicates one requirement of relevance. Intuitively, the model checks if each of the requirements is satisfied by checking spans in $\mathcal{D}$. Assuming that the model performs such matching process, *alignment* explanations provides more sensible description of model behavior compared to token- or word-level explanations [15].

To evaluate alignment rationales for relevance ranking, we need metrics that capture the degree to which the rationales extracted by an explanation model are in fact contributed to the model prediction. Our goal is to define metrics for evaluating the *faithfulness* of alignment rationales. Once the evaluation metrics are established, they can be used as bases for alignment generation methods, by optimizing the proposed quality metrics via black-box optimization [6] or gradient-based methods [9].

Problem Definition. Assume an alignment (qt, dt), where $qt \in \mathcal{Q}$ and $dt \in \mathcal{D}$, is given by an explanation model when $f(q,d) \geq \theta_r$, i.e., the document d is predicted to be relevant to the query q by the model f. The goal is to measure the faithfulness of this alignment to the behavior of model f.

4 EVALUATION METRICS

We use the two criteria sufficiency and necessity [1] in our metrics. **Sufficiency** measures whether a rationale is sufficient for a model prediction by comparing the model output for the full input to its output for the input built from the rationale. **Necessity** measures whether a rationale captures only the necessary information by comparing the model output when the rationale is removed.

We first introduce how these metrics can be used to check alignment-independent rationale for document relevance, and show why they are not suitable for evaluating the faithfulness of alignment rationale explanations. We then propose a new set of metrics.

4.1 Alignment-Independent Metrics

Given that $R(q,d) = 1$, a document span dt provides sufficient relevant information if $R(q,dt) = 1$, where the input dt to the model means the document content except the span dt is deleted or masked. Formally, this metric can be defined based on real-valued output (continuous score) or based on binary relevance labels of the model as follows.

$$\text{AI.Suff}(q,d,dt) = -[f(q,d) - f(q,dt)], \tag{1}$$

$$\text{AI.Suff}_\text{b}(q,d,dt) = \mathbb{1}[\theta_r \leq f(q,dt)], \tag{2}$$

[1] Code for reproducing experiments will be made available at https://github.com/youngwoo-umass/alignment_rationale
[2] Based on binary-necessity category.

$\mathbb{1}[.]$ is an indicator function, returning a value of one when its condition is satisfied. We use $\cdot_b$ (such as AI.Suff$_b$) to denote the metrics based on binary outputs. Similar notations are used for the following metrics too. The negative sign is added to make a higher value (closer to zero) of AI.Suff indicate a higher quality of rationale.

The sufficiency metric prefers longer spans of documents as explanations. For example, in the extreme case of selecting the entire document as an explanation, the metric will have the highest value. To address this issue, we propose a modification of the *necessity* metric [1] for relevance ranking. Let $d \setminus dt$ denotes a text acquired by removing the span dt from document d. Our *necessity* metric considers the span dt as having only the necessary relevant information and being compact if $R(q, d \setminus \hat{dt}) = 0$ for all non-empty $\hat{dt} \subseteq dt$. This metric penalizes long explanations containing non-relevant information.

$$\text{AI.Ness}(q, d, dt) = f(q, d) - \text{avg}_{\hat{dt} \subseteq dt} f(q, d \setminus \hat{dt}) \tag{3}$$

$$\text{AI.Ness}_b(q, d, dt) = \mathbb{1}[f(q, d \setminus \hat{dt}) < \theta_r] \tag{4}$$

It is computationally expensive to compute the outputs of a deep neural model for several subsets of each candidate span. Therefore, we randomly sampled 10%, 20%, ..., 100% of dt as $\hat{dt}$ and averaged the model predictions for these subsets.

While these metrics evaluate whether span dt has contributed to the model's relevance prediction, they do not evaluate whether it has been aligned with query span qt or not. These definitions are all based on deletion perturbations of the instance (q, d) to be explained. We thus start by extending these metrics for evaluation of the alignment faithfulness using deletion perturbations.

4.2 Deletion-based Metrics

To evaluate alignment rationales, we consider simultaneous perturbations of the query and document in the instance to be explained. One would intuitively expect that if a document is relevant to a query, it is also relevant to any span of the query. Specifically, when $R(q, d) = 1$, the expectation is to get $R(qt, d) = 1$. If this assumption is satisfied by the model, we can perturb the document to extract the span dt that affects prediction $R(qt, d) = 1$ and validate the influence of alignment (qt, dt) in the model prediction $f(q, d)$. Given the condition $R(qt, d) = 1$, the evaluation metrics are then formally defined as follows.

Sufficiency. Span dt provides sufficient relevant information for span qt if $R(qt, dt) = 1$. This metric is referred to as D.Suff.

$$\text{D.Suff}(q, d, qt, dt) = -[f(qt, d) - f(qt, dt)] \tag{5}$$

$$\text{D.Suff}_b(q, d, qt, dt) = \mathbb{1}[\theta_r \leq f(qt, dt)] \tag{6}$$

Necessity. Span dt contains only the necessary relevant information for span qt if $R(qt, d \setminus \hat{dt}) = 0$ for all non-empty $\hat{dt} \subseteq dt$.

$$\text{D.Ness}(q, d, qt, dt) = f(qt, d) - f(qt, d \setminus \hat{dt}) \tag{7}$$

$$\text{D.Ness}_b(q, d, qt, dt) = \mathbb{1}[f(qt, d \setminus \hat{dt}) < \theta_r] \tag{8}$$

4.3 Substitution-based Metrics

Deletion-based metrics rely on the implicit assumption that $R(qt, d) = 1$ when $R(q, d) = 1$. However, this assumption frequently fails. For example, the ranker [3] that is used in our experiments predicted that the document in Figure 1 is relevant to the query "Where is SIGIR 2022", but it is not relevant to the query "Where is". To address this issue, we propose to substitute the query parts other than qt instead of deleting them.

We introduce a new query $qt \cup w$, which is built by substituting spans of $q \setminus qt$ with spans w. For the example query "Where is SIGIR 2022", qt can be "Where is". Deletion-based metrics use the model prediction for query "Where is" to compute faithfulness. Instead, we substitute "SIGIR 2022" with w ="CIKM 2022", and probe the model with the new query "Where is CIKM 2022". As spans w are newly introduced to query, it is likely that the document does not contain any information about w. Thus, we also add w to span dt of the document so that the w part of the new query has exact match in the document. Substitution allows to more accurately measure if the qt part of the query is satisfied by the document span dt.

Sufficiency. Span dt provides sufficient relevant information for span qt if $f(qt \cup w, dt \cup w) = 1$.

$$\begin{aligned} \text{S.Suff}(q, d, qt, dt) &= -[f(qt \cup w, d \cup w) - f(qt \cup w, dt \cup w)] \\ \text{S.Suff}_b(q, d, qt, dt) &= \mathbb{1}[\theta_r \leq f(qt \cup w, dt \cup w)] \end{aligned} \tag{9}$$

Necessity. Span dt contains only the necessary relevant information for span qt if $R(qt + w, dt \setminus \hat{dt} \cup w) = 0$ for all non-empty $\hat{dt} \subseteq dt$.

$$\begin{aligned} \text{S.Ness}(q, d, qt, dt) &= f(qt \cup w, d) - f(qt \cup w, dt \setminus \hat{dt} \cup w) \\ \text{S.Ness}_b(q, d, qt, dt) &= \mathbb{1}[\exists w \text{ s.t. } f(qt \cup w, dt \setminus \hat{dt} \cup w) < \theta_r] \end{aligned} \tag{10}$$

Substitution candidates. When substitution spans have the same syntactic role and similar semantic category as $q \setminus qt$, the new query is more comparable to the original query. However, we found that even without such complex selection of w, $qt \cup w$ can provide more reliable estimate of model behavior. To get substitution candidates, we first collect all term-level n-grams of the target retrieval collection for values of n ranging from 1 to 4. A span w from the obtained n-grams will be used for the computation of the substitution-based metrics if $f(qt \cup w, w) = 0$. This condition allows to prune the large candidate space and to make sure that selected spans are not specific enough that their matching alone is enough for relevance prediction by the model for $(qt \cup w, d \cup w)$. If no span w satisfies the condition, the lower bound score is assigned to S.Suff and S.Ness.

5 EXPERIMENTS

The experiments demonstrate how the metrics proposed in Section 4 are different. We are especially interested in comparing deletion-based versus substitution-based metrics and binary versus continuous metrics.

Dataset. We use the BERT-based document ranker as our target function f to be explained [3]. We trained the model with MS-MARCO document ranking dataset [2], and perform alignment evaluation on the dev split. The ranker is trained with the cross-entropy loss, thus can be considered as a binary classifier. The trained ranker showed NDCG@10 of 0.625 on TREC Deep Learning Track 2019 [2], which matches the performance reported by the similar models [2].

Table 1: Preferences and accuracy of different metrics on two alignment methods: exact match (EM) and random. The cases where the difference between the deletion and substitution metrics are statistically significant ($p < 0.01$) are denoted with *. The numbers in bold (substitutions) are the ones that we consider better compared to the corresponding deletion based version.

Metrics			Relative preference			Accuracy	
			EM	Random	Equal	EM	Random
Continuous	Attention Mask		0.50	0.50	0.00	0.86	0.86
Binary	Necessity	Deletion ($D.Ness_b$)	0.13*	0.00*	0.87*	0.98*	0.87*
		Substitution ($S.Ness_b$)	0.66*	0.02*	**0.32***	0.92*	0.33*
	Sufficiency	Deletion ($D.Suff_b$)	0.78*	0.00	0.22*	0.83*	0.06*
		Substitution ($S.Suff_b$)	0.81*	0.01	**0.18***	0.97*	0.16*
Continuous	Necessity	Deletion (D.Ness)	0.85	0.15	0.00	0.99*	0.87*
		Substitution (S.Ness)	0.86	0.14	0.00	0.94*	0.34*
	Sufficiency	Deletion (D.Suff)	0.97	0.03	0.00	0.83*	0.06*
		Substitution (S.Suff)	0.97	0.03	0.00	0.97*	0.16*

Table 2: Accuracy of exact match alignments for different units of query-side targets (qt).

		word	low-idf spans	high-idf spans
Attention Mask		0.86	0.62	0.95
Necessity	Deletion	0.98	0.88	0.64
	Substitution	0.92	0.90	0.73
Sufficiency	Deletion	0.83	0.21	0.74
	Substitution	0.97	0.75	0.91

Our main evaluation set consists of 3,176 cases where each case consists of a unique triple (query, text, query-side target qt). These cases are obtained by selecting 50 queries and the documents that are predicted to be relevant to them. We split documents by sentences, and filtered sentences that are predicted to be relevant to the query when they are fed individually. In the main evaluation setting, individual words are used as a query-side target.

Alignments. We analyze the behavior of each metric on two alignment methods: exact and random matches. The exact match alignment is built by selecting any word in the document that overlaps with the words of the query-side target qt. The overlap is compared in sub-word level. Random alignments are built by randomly selecting document tokens. Random alignments are controlled to have the same number of tokens as the exact-match alignments.

We assume that the exact match alignments are better than random *on average*. Thus, we can expect that an ideal metric prefers exact-match over random alignments. This does not imply that the ideal metric should prefer exact-match over random for every case since it is possible that in some cases random alignments may be better than the exact-match alignments. When measuring relative preferences of alignments by evaluation metrics, the cases where no exact match exists are excluded.

We also compare our metrics with another evaluation metric for alignments based on attention masks [9]. This metric drops attention flows between the two segments (query and document), except the token pairs that are predicted to be aligned. Section 2 provides more details about this metric. For the attention-mask metric, the binary version is not applicable because changing the attention mask results in a change of the model score by a small magnitude only, which does not flip the classification label (always relevant). Following Jiang et al. [9], the absolute difference of logistic scores are used to compute the metric.

Results. Table 1 shows the results of the various metrics on the two types of alignments. Relative preference shows how often exact-match or random alignment is preferred over the other by a metric. We removed the cases where exact match (EM) and Random had the same alignment prediction. Thus, the equal column of the table indicates the rates that randomly-aligned and exact-match tokens get the same preference by an evaluation metric, while the tokens are different. Accuracy indicates the rate that the score given by a metric is over the $\theta_r = 0.5$ (in case of attention mask, lower than θ_r).

First, we observe that the attention-mask metric does not behave as expected. It prefers random alignments in almost half of the cases, which implies that this metric is capturing something different than the alignment rationales for relevance ranking. We investigated these cases to find out which tokens appear when the random alignment is preferred. The tokens for some special characters such as "." or "?" appear more often in the preferred cases than their average frequencies. This implies that if an alignment contains "." or "?", it is more likely to be preferred over the exact match by the attention-mask metric compared to when the alignment contains other random tokens. One potential reason can be that the BERT-based ranker is using the tokens for these special characters to combine information, thus the matching tokens (such as common words in the query and document) are compared via these tokens. Another possibility can be that removing attentions between these tokens breaks the score calculation even if they do not play a role in the matching process.

Second, we observe that the `substitution`-based metrics have a lower rate of equal decisions compared to their corresponding `deletion`-based metrics. We conclude that the high equal rate of necessity-`deletion` metric indicates a clear failure of the metric. First, the cases with the same alignment were removed, thus compared alignments are always different. Second, the dataset is known to have many exact match terms between the queries and documents, thus a certain portion of exact-match alignments should be

considered better than the random ones. Thus, we conclude that the substitution-based metrics have advantages over the deletion-based metrics. We approximate the "coverage" of a metric as the portion of the data that the metric makes two different decisions on two different alignments. In case of binary-necessity, the deletion-based metric has coverage of 13% (13% + 0%) and substitution-based metric has coverage of 68% (66% + 2%).

We believe that the high accuracy of the necessity metrics is probably resulted from the perturbed queries (qt or $qt + w$) not being comparable to their corresponding original queries, thus yielding non-relevant predictions for all perturbations.

Next, we compare the binary metrics against their continuous versions. With the continuous metrics, the equal rate decreased to near zero. A large portion of the equal cases by the binary metrics are classified as preference to exact matches by the continuous metrics. From this trend, we expect that continuous metrics, that are sensitive to small differences in model scores, could be capable of preferring better alignments. Reduction in equal cases of the Necessitymetric is mostly observed for cases that $f(qt, d)$ is near zero, $f(qt, d \setminus \hat{dt})$ for exact match dt is also near zero and is lower than $f(qt, d \setminus \hat{dt})$ for randomly aligned dt.

Query-side target Finally, we compare the evaluation metrics when different segmentation units of queries are used for explanation, i.e., different query targets qt. We built two datasets "high-idf spans" and "low-idf spans". The original dataset consisting of individual words as qt is called "word". For each query, we identified the query terms whose idf (inverse document frequency) values exceed a predefined threshold value. We select a continuous span of the query that covers these high-idf terms. These high-idf spans compose the "high-idf spans" dataset. The remaining low-idf terms, which can be at most two continuous segments per query, compose "low-idf spans". For example, "Where is" constitutes the low-idf span and "SIGIR 2022" constitutes the high-idf span for the example query "Where is SIGIR 2022". We expect the high-idf spans, such as entity names, to have more exact matches, because they are considered to be more important in determining relevance. In contrast, low-idf spans contain frequent words such as wh-words or stopwords (e.g., "where is"). Thus, exact match alignments would be less effective for low-idf spans.

Table 2 shows the accuracy of the exact match alignments on three span units: word, high-idf spans, and low-idf spans. Only scores for binary versions of metrics are reported as they are nearly identical to their corresponding continuous versions. The accuracy of low-idf spans and for high-idf spans is considerably lower than that of words. We attribute this to the fact that 'word' test set is too favorable to exact match, as it only considers cases when exact match is found. However, in these datasets with longer spans, some query terms in query-side target (qt) may not appear in the document, which would lead to a lower performance of exact-match alignments. We can also observe that the difference between deletion-based metrics and substitution-based metrics gets larger in cases of sufficiency groups on low-idf spans.

6 CONCLUSION

This paper studies how the perturbation-based metrics can be used to evaluate alignment rationales for black-box document ranking models. The concepts of necessity and sufficiency are defined and applied to simultaneous perturbations of the query and document pair. Deletion-based metrics and substitution-based metrics are defined for each of the two concepts. The experiments show the characteristics of the metrics and demonstrate that substitution-based metrics are more successful than the deletion-based ones in preferring higher-quality alignments.

7 ACKNOWLEDGMENT

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant number 1813662. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and Characterizing Human Rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 9294–9307. https://doi.org/10.18653/v1/2020.emnlp-main.747

[2] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. [n.d.]. OVERVIEW OF THE TREC 2019 DEEP LEARNING TRACK. ([n. d.]).

[3] Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 985–988.

[4] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4443–4458. https://doi.org/10.18653/v1/2020.acl-main.408

[5] Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. 2019. A Study on the Interpretability of Neural Retrieval Models Using DeepSHAP. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. 1005–1008.

[6] Peter Hase, Harry Xie, and Mohit Bansal. 2021. The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations. *Advances in Neural Information Processing Systems* 34 (2021).

[7] Marti A Hearst. 1995. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 59–66.

[8] Orland Hoeber and Xue Dong Yang. 2006. The Visual Exploration ofWeb Search Results Using HotMap. *Tenth International Conference on Information Visualisation (IV'06)* (2006), 157–165.

[9] Zhongtao Jiang, Yuanzhe Zhang, Zhao Yang, Jun Zhao, and Kang Liu. 2021. Alignment Rationale for Natural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5372–5387.

[10] Youngwoo Kim, Myungha Jang, and James Allan. 2020. Explaining text matching on neural natural language inference. *ACM Transactions on Information Systems (TOIS)* 38, 4 (2020), 1–23.

[11] Youngwoo Kim, Razieh Rahimi, Hamed Bonab, and James Allan. 2021. *Query-Driven Segment Selection for Ranking Long Documents*. Association for Computing Machinery, New York, NY, USA, 3147–3151. https://doi.org/10.1145/3459637.3482101

[12] 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 *CoCo@ NIPS*.

[13] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. *arXiv preprint arXiv:1904.07531* (2019).

[14] Razieh Rahimi, Youngwoo Kim, Hamed Zamani, and James Allan. 2021. Explaining Documents' Relevance to Search Queries. *arXiv preprint arXiv:2111.01314* (2021).

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[16] Procheta Sen, Debasis Ganguly, Manisha Verma, and Gareth J.F. Jones. 2020. *The Curious Case of IR Explainability: Explaining Document Scores within and across Ranking Models.* 2069–2072.

[17] Jaspreet Singh and Avishek Anand. 2018. Posthoc Interpretability of Learning to Rank Models using Secondary Training Data. In *Workshop on ExplainAble Recommendation and Search (EARS 2018) at SIGIR 2018.*

[18] Jaspreet Singh and Avishek Anand. 2019. Exs: Explainable search using local model agnostic interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 770–773.

[19] Jaspreet Singh, Megha Khosla, Wang Zhenye, and Avishek Anand. 2021. Extracting per Query Valid Explanations for Blackbox Learning-to-Rank Models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 203–210. https://doi.org/10.1145/3471158.3472241

[20] Manisha Verma and Debasis Ganguly. 2019. LIRME: Locally Interpretable Ranking Model Explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. 1281–1284.

[21] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. *Pretrained Transformers for Text Ranking: BERT and Beyond.* Association for Computing Machinery, New York, NY, USA, 1154–1156. https://doi.org/10.1145/3437963.3441667

[22] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. *An Analysis of BERT in Document Ranking.* Association for Computing Machinery, New York, NY, USA, 1941–1944. https://doi.org/10.1145/3397271.3401325

[23] Honglei Zhuang, Xuanhui Wang, Michael Bendersky, Alexander Grushetsky, Yonghui Wu, Petr Mitrichev, Ethan Sterling, Nathan Bell, Walker Ravina, and Hai Qian. 2021. Interpretable Ranking with Generalized Additive Models. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 499–507.