THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

DREQ: Document Re-Ranking Using Entity-based Query Understanding

OPEN ACCESS

DREQ: Document Re-Ranking Using Entity-based Query Understanding

Shubham Chatterjee[1][0000−0002−6729−1346], Iain Mackie[2], and
Jeff Dalton[3][0000−0003−2422−8651]

[1] University of Edinburgh, UK `shubham.chatterjee@ed.ac.uk`
[2] University of Glasgow, UK `i.mackie.1@research.gla.ac.uk`
[3] University of Edinburgh, UK `jeff.dalton@ed.ac.uk`

Abstract. While entity-oriented neural IR models have advanced significantly, they often overlook a key nuance: the varying degrees of influence individual entities within a document have on its overall relevance. Addressing this gap, we present `DREQ`, an entity-oriented dense document re-ranking model. Uniquely, we emphasize the query-relevant entities within a document's representation while simultaneously attenuating the less relevant ones, thus obtaining a query-specific entity-centric document representation. We then combine this entity-centric document representation with the text-centric representation of the document to obtain a "hybrid" representation of the document. We learn a relevance score for the document using this hybrid representation. Using four large-scale benchmarks, we show that `DREQ` outperforms state-of-the-art neural and non-neural re-ranking methods, highlighting the effectiveness of our entity-oriented representation approach.

1 Introduction

During the last decade, the emergence of large-scale Knowledge Graphs (KGs) has motivated the development of entity-oriented search systems. Entities, with their rich semantic information, help bridge the gap between unstructured text and structured knowledge. Prior research [16,61,62,36,55] underscores the significance of entities in feature-based retrieval systems. Their effectiveness within neural IR models has also been demonstrated, with the Entity-Duet Ranking Model (EDRM) from Liu et al. [38] standing out as a pioneering effort. This model synergizes the word-entity duet framework [62] with the strengths of neural networks and KGs. More recently, Tran and Yates [59] introduced an approach that clusters entities within documents, offering multiple "views" or perspectives to enhance the understanding of various document facets.

Yet, amidst these advancements, current models often overlook a crucial aspect: not all entities within a document contribute equally to its relevance. For instance, given the query "Black Bear Attacks" and the document in Figure 1, while the entity *hand axe* may have peripheral relevance to the given query due to its defensive use during animal encounters, the entity *National Guard* is

likely non-relevant. Conversely, the entity *Parks Highway*, known for bear sightings, may hold greater significance. In the "retrieve-then-rerank" paradigm, a common approach in neural IR, initial retrieval fetches a broad set of candidate documents. For the subsequent re-ranking phase, though, the differential relevance of individual entities becomes particularly critical as the model must sift through the candidates and rank them with high precision. Furthermore, prior work [17,62,38,59] often produce query-agnostic document representations which fail to resonate with the specific nuances and requirements of the query. Our proposition emphasizes a more refined approach: to enhance re-ranking accuracy, entities should be weighted based on their query relevance, ensuring that the document's representation is both influenced by the most relevant entities and tailored to the query's nuances.

Against this backdrop, in this work[1], we introduce ***Document Re****-ranking using **E***ntity-based **Q***uery Understanding* (DREQ), an entity-oriented neural re-ranking model that extends the conventional "retrieve-then-rerank" paradigm by introducing an innovative intermediate step. Given a query Q and a candidate set of documents $\mathcal{D}$ retrieved using an initial retrieval method (e.g., BM25), we want to *re-rank* these candidates to order them by their relevance to the query Q. While prior approaches predominantly utilize entities to gain a fine-grained understanding of documents, our method uniquely perceives entities within candidate documents as overarching concepts essential for a comprehensive understanding of the query. We identify and prioritize entities that align closely with the query. To this end, we emphasize the embeddings of the relevant entities and concurrently attenuate the less relevant ones within the document's representation, thus obtaining a refined, *query-specific* and *entity-centric* document representation. Recognizing that the raw text of a document captures the document's overarching semantic context, we also derive a broader text-centric representation of the document, complementing the focused entity-centric perspective. We meld the entity-centric and text-centric representations of the document, obtaining a "hybrid" document representation that imbibes insights about what the query specifically seeks. Using this hybrid representation, we learn a fine-grained "interaction vector" that imbibes the differences, commonalities, and subtle relationships between the query and document. We then use this vector to learn a relevance score for the document. We term our methodology "retrieve-harness-rerank"[2], emphasizing our process of harnessing the information from entities present in candidate documents to re-rank the candidates.

Contributions. We make the the following contributions through this work:

1. We introduce DREQ, an entity-oriented re-ranking model that enriches a document's representation with *query-specific* entity knowledge for nuanced relevance matching. We achieve new state-of-the-art results on four major document ranking test collections.

---

[1] **Code and data: https://github.com/shubham526/ECIR2024-DREQ.**

[2] Term coined by Dr. Laura Dietz at the SIGIR 2023 tutorial (https://github.com/laura-dietz/neurosymbolic-representations-for-IR).

2. We introduce a hybrid representation learning mechanism, blending entity-centric and text-centric representations. This hybrid representation captures a document's broad context and query-specific relevance, enhancing the precision of re-ranking.

3. We show that the effectiveness of re-ranking is significantly enhanced by meticulously selecting and assigning appropriate weights to entities within a document.

2   Related Work

Entity-Oriented Search. Initial attempts at entity-oriented search primarily used entities for query expansion, one notable example being Entity-Query Feature Expansion [11] (EQFE) model which used entity links within documents for query expansion. Entities later became a latent layer [17,36,61] in document and query representations, forming a high-dimensional entity space that improved retrieval by revealing hidden semantics. Research progressed to treat entities as explicit elements in retrieval models, coexisting with term-based approaches. Methods like entity-based language models [55] and semantically-driven models [16] rank documents by their semantic relation to the query. A prominent line of research [66,62,64,61,65] introduced a dual-layered approach, combining a "bag-of-entities" with the traditional "bag-of-terms" to improve document retrieval.

Neural IR. Recently, deep learning has transformed text ranking, removing the need for handcrafted features and fostering semantic matching. Before BERT [12], models either created vector representations [25,57,43,44] for queries and documents or built similarity matrices to capture term interactions [22,63,26,27,10]. The advent of BERT and its derivatives [37,7,24,72,28] in a retrieve-then-rerank [49,50] framework ushered in models like Birch [1] and BERT-MaxP [8], emphasizing sentence or passage relevance. Subsequent models, such as CEDR [39] and PARADE [35], utilized BERT's contextual embeddings, with the latter aggregating passage representations. Concurrently, ERNIE [72] enhanced BERT with knowledge graphs and vast text corpora.

Meanwhile, bi-encoders emerged as a foundational mechanism in text ranking, employing distinct encoders to derive query and document vectors. For example, DPR [30] employs BERT's `[CLS]` token for both, gauging similarity through their inner product. The choice of negative examples is crucial: DPR favors BM25-retrieved passages while ANCE [68] leans on ANN techniques. Addressing bi-encoders' challenges with term-level interactions, research has delved into multi-vector text representation, exemplified by ColBERT's [31] "late interaction" using token-level embeddings.

The evolution of pre-trained sequence-to-sequence models like T5 [53] has led to models such as MonoT5 [51] and RankT5 [76]: The former assesses document relevance via probability assignment to "true" labels during decoding whereas the latter provides direct ranking scores for query-document pairs.

Query/Document Expansion. Recent efforts in this direction have been driven by embeddings. Solutions such as DeepCT [9] and docT5query [51] use
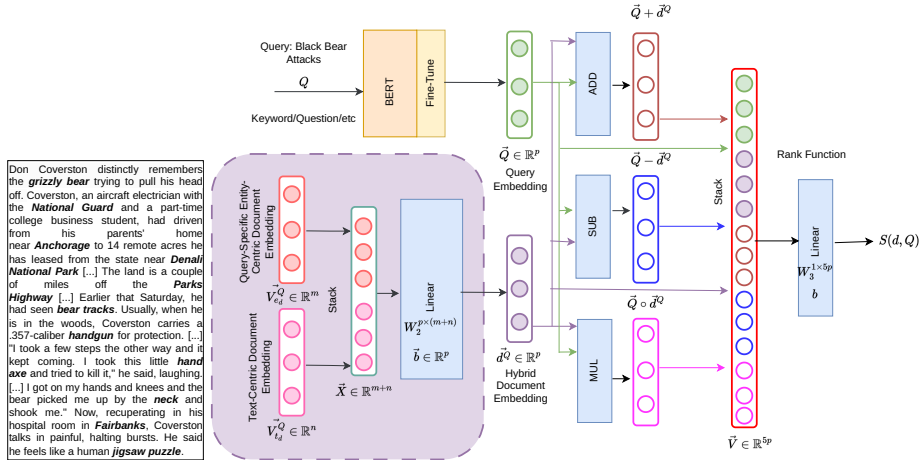
Fig. 1: Our proposed system `DREQ` uses a hybrid document embedding learnt using (1) the query-specific entity-centric embedding, and (2) text embedding of the document to learn the document score.

transformers to enrich traditional PRF models by highlighting significant terms in documents. Simultaneously, methods like Neural PRF [34] and BERT-QE [73] have adopted neural ranking models to measure document similarity with feedback documents. CEQE [45] has further enhanced this approach by using BERT to create contextualized representations and select expansion terms based on similarity in the embedding space. While ColBERT-PRF [60] directly utilizes BERT embeddings for retrieval without further training, ANCE-PRF [71] requires additional retraining of the query encoder using PRF information

Entity Ranking. Early methods include using MRFs to deal with the joint distribution of entity terms from semi-structured data [42,74,48,23,54], utilizing types [29,2,18] and relations [58,6] from a KG, and using Learning-To-Rank methods [56,21,13,4] to rank entities using a diverse set of features. Recent advancements have been significantly driven by neural models using advanced techniques such as autoregressive entity ranking [3], integrating BERT-based entity rankings with other features [5], and enhancing BERT with Wikipedia2Vec [69] embeddings [20]. Concurrently, the emergence of graph embedding-based models [19,47] has enriched entity ranking approaches by utilizing joint embedding of entities and words within the same vector space.

3   Approach: DREQ

Overarching Idea. Our work is anchored in the intuition that documents contain interconnected entities which provide a distilled understanding of the document's main content, and serve as broad concepts important for understanding the query. For example, in Figure 1, given the query "Black Bear Attacks", the presence of specific geographical entities such as *Parks Highway* and *Fairbanks*

narrows down the context to the Alaskan wilderness, a region known for its bear population. Additionally, the entity *Hand axe* sheds light on human defensive measures during such encounters. Together, these entities reinforce the document's fit for narratives around bear encounters. Even though the central narrative revolves around a grizzly, the web of interconnected entities creates an ecosystem of information that can be invaluable in assessing relevance to the broader theme of bear attacks.

Entity Ranking. Nonetheless, the importance of these entities for determining the document's relevance is query-dependent. For example, in the context of "Black Bear Attacks", the entity *Parks Highway* is highly important as it specifies a geographical context known for its black bear population. However, for a broader query like "Human-animal conflict in North America", the entity is less significant as the theme encompasses a vast range of conflicts, animals, and locations, making the specific location of *Parks Highway* just one of many possible locales of interest and less indicative of the document's relevance.

To address this issue, we pool all entities from all candidate documents $d \in \mathcal{D}$ to obtain a candidate set of entities $\mathcal{E}$ for the query Q. Following previous work [46,15,14], we transfer the relevance labels from documents to entities in the document based on the assumption that a relevant document contains relevant entities. Using this entity ground truth, we train a separate entity ranking model to rank entities $e \in \mathcal{E}$. Specifically, we follow previous work [41,67,38,70,5] and leverage the Knowledge Base (DBpedia [33]) description t_e of entity e to learn an embedding $\mathbf{e} \in \mathbb{R}^k$ of the entity using BERT. The input to BERT is a sequence of query tokens $\tau^q \in Q$ and description tokens $\tau^e \in t_e$, separated by the special token [SEP], and preceded by the special token [CLS]. We use the k-dimensional embedding of the [CLS] token from the last hidden layer of BERT as the embedding $\mathbf{e}$ of an entity $e \in \mathcal{E}$. To derive a rank score for this entity in relation to the query Q, we learn a linear projection $\mathbb{R}^k \to \mathbb{R}$ using the embedding $\mathbf{e}$ and a weight matrix $W_1^{1 \times k}$. This scoring function $\mathrm{S}(e, Q)$ is formulated as: $\mathrm{S}(e, Q) = W_1 \cdot \mathbf{e} + b$, where b is a scalar bias term.

Document Representation. Entities such as *Parks Highway*, *Neck*, and *Hand axe* in Figure 1 collectively suggest a bear encounter in Alaska. When related to the query "Black Bear Attacks", the combined narrative from the entities implies the document is probably relevant as it aligns with the query's essence. Based on this idea, and acknowledging that the significance of these entities for determining a document's relevance varies with the query, we learn a ***query-specific entity-centric*** representation $\mathbf{V}_{e_d}^Q \in \mathbb{R}^m$ of a candidate document $d \in \mathcal{D}$. Specifically, we first represent each entity within a document via its embedding from Wikipedia2Vec [69] due to its ability to capture the relationships and deeper contexts between entities. The document representation $\mathbf{V}_{e_d}^Q$ is a weighted sum of embeddings of entities in the document:

$$\mathbf{V}_{e_d}^Q = \sum_{e \in d} w_e \cdot \mathbf{e} \tag{1}$$

where the weight w_e for each entity is query-specific, determined by the rank score of the entity:

$$w_e = \frac{S(e, Q)}{\sum_{e' \in d} S(e', Q)}$$

Note that $S(e, Q)$ are normalized using the `softmax` function. Consequently, the weights w_e can be interpreted as the *probability* of the entity's relevance to the query. Intuitively, using these weights would prioritize query-relevant entities while concurrently down-weighting those of lesser relevance within the document.

While encapsulating key entities helps capture a document's query-specific semantic essence, it's crucial to also consider the broader textual context to capture nuances missed by solely focusing on entities. Hence, we also obtain a text-centric representation $\mathbf{V}_{t_d}^Q \in \mathbb{R}^n$ of a candidate document $d \in \mathcal{D}$ using BERT. Following previous work [39,35], we first segment the document into passages using a sliding window of M sentences with a stride of S sentences over the document. We represent each passage in the document via the embedding of the `[CLS]` token from BERT. To obtain the document representation $\mathbf{V}_{t_d}^Q$, we average the passage embeddings. We combine the text and entity-centric document representations to learn a single ***hybrid*** representation $\mathbf{d}^Q \in \mathbb{R}^p$ via a linear projection $\mathbb{R}^{m+n} \to \mathbb{R}^p$ as follows:

$$\mathbf{d}^Q = W_2 \cdot [\mathbf{V}_{t_d}^Q; \mathbf{V}_{e_d}^Q] + \mathbf{b}$$

where $\mathbf{X} \in \mathbb{R}^{m+n}$ is the concatenated embeddings $[\mathbf{V}_{t_d}^Q; \mathbf{V}_{e_d}^Q]$, $W_2^{p \times (m+n)}$ is a weight matrix, and $\mathbf{b} \in \mathbb{R}^p$ is a bias vector. Our intuition is that this hybrid embedding $\mathbf{d}^Q$ of a candidate document $d \in \mathcal{D}$ encapsulates both the granular insights from the entities and the overarching narrative provided by the document's text, thereby ensuring a more accurate search.

Document Ranking. To learn the document scoring function $\mathrm{S}(d, Q)$, we first learn several fine-grained interactions between the query embedding $\mathbf{Q} \in \mathbb{R}^p$ (obtained via the embedding of the `[CLS]` token from BERT) and the hybrid document embedding $\mathbf{d}^Q$ as follows: An additive interaction $\mathbf{V}_{\mathbf{add}}^{\mathbf{d}, \mathbf{Q}} = \mathbf{Q} + \mathbf{d}^{\mathbf{Q}}$, subtractive interaction $\mathbf{V}_{\mathbf{sub}}^{\mathbf{d}, \mathbf{Q}} = \mathbf{Q} - \mathbf{d}^{\mathbf{Q}}$, and multiplicative (Hadamard product) interaction $\mathbf{V}_{\mathbf{mul}}^{\mathbf{d}, \mathbf{Q}} = \mathbf{Q} \circ \mathbf{d}^{\mathbf{Q}}$. We then learn the scoring function $\mathrm{S}(d, Q)$ through a linear projection $\mathbb{R}^{5p} \to \mathbb{R}$ as follows:

$$\mathrm{S}(d, Q) = W_3 \cdot \mathbf{V} + b$$

where $W_3^{1 \times 5p}$ is a weight matrix, $\mathbf{V} \in \mathbb{R}^{5p}$ is a vector representing the concatenated embeddings $[\mathbf{Q}; \mathbf{d}^{\mathbf{Q}}; \mathbf{V}_{\mathbf{add}}^{\mathbf{d}, \mathbf{Q}}; \mathbf{V}_{\mathbf{sub}}^{\mathbf{d}, \mathbf{Q}}; \mathbf{V}_{\mathbf{mul}}^{\mathbf{d}, \mathbf{Q}}]$, and b is a scalar bias. We hypothesize that by subtracting, adding, and multiplying the query and hybrid document embeddings, we can explore different aspects of how the query relates to the document. Subtraction might show the differences between them, which can help point out any areas that don't match. On the other hand, addition might reveal what they have in common, highlighting overlapping themes.

Multiplying the embeddings goes a step further and might reveal subtle relationships between specific parts of the query and document. By merging these interactions with the original query and document embeddings, our intuition is that the model would garner a rich, composite insight into the document's depth and breadth.

End-To-End Training. We train both our entity and document ranking models using the binary cross-entropy loss below:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\Big(y_i \cdot \log(p(\hat{y}_i)) + (1 - y_i) \cdot (1 - \log(p(\hat{y}_i)))\Big)$$

where y_i is the label and $p(\hat{y}_i)$ is the predicted probability of the entity/document being relevant. This is analogous to pointwise learning-to-rank where each of N query-entity/query-document pairs is independently classified into relevant or non-relevant. The entire model is optimized end-to-end using back-propagation. The weight matrices, namely W_1 for the entity ranking model and (W_2, W_3) for the document ranking model, are learnt concurrently with the fine-tuning of the initial embeddings. Moreover, the Wikipedia2Vec embeddings undergo end-to-end fine-tuning specifically within the document ranking model, ensuring that these entity embeddings are tailored to enhance document ranking effectiveness.

4   Experimental Methodology

4.1   Datasets

CODEC. CODEC [40] is a benchmark specifically designed for complex research topics in social sciences. The benchmark includes 42 topics, and new entity linked corpus with 729,824 documents with focused content across finance, history, and politics. The corpus contains an average of 159 entities per document. It provides expert judgments on 6,186 documents derived from diverse automatic and manual runs.

TREC Robust 2004. The TREC Robust 2004 track focuses on poorly performing topics. The track provides 250 topics with short "titles" and longer "descriptions"; we report results on both title and description queries. The collection consists of 528,024 documents (containing an average of 116 entities per document) taken from TREC disks 4 and 5 excluding the Congressional Record. The track provides 311,409 graded relevance judgments for evaluation.

TREC News 2021. The TREC News track focuses on search tasks within the news domain. We focus on the background linking task, which involves retrieving news articles that provide relevant context or background information for a given news story. There are 51 topics, each with a title, description, and narrative; in this work, we use all three fields for query formulation. The track uses the TREC Washington Post (v4) collection, encompassing 728,626 documents containing 131 entities per document on average. The track provides 12,908 graded relevance assessments for evaluation.

TREC Core 2018. The TREC Core track offers 50 topics, each consisting of titles, descriptions, and narratives. For this work, we utilize all three components. The track uses the TREC Washington Post (v2) collection, encompassing 595,037 news articles and blog posts and containing 123 entities per document on average. 26,233 graded relevance judgements are available.

4.2   Evaluation Paradigm

Candidate Ranking. We retrieve a candidate set of 1000 documents per query using BM25+RM3 (Pyserini default). The Recall@1000 of this candidate set for each dataset is as follows: (1) CODEC: 0.82, (2) Robust04 (title): 0.77, (3) Robust04 (desc): 0.75, (4) News 2021: 0.94, (5) Core 2018: 0.75. Other metrics shown in Tables 1 and 2

Evaluation Metrics. (1) Precision at $k = 20$, (2) Normalized Discounted Cumulative Gain (nDCG) at $k = 20$, and (3) Mean Average Precision (MAP). We conduct significance testing using paired-t-tests.

Entity Linking. Our work relies on an entity linked corpus. While any entity linking system may be used, in this work, we use WAT [52].

Train and Test Data. As positive examples during training, we use documents that are assessed as relevant in the ground truth provided with the dataset. Following the standard [30], for negative examples, we use documents from the candidate ranking (BM25+RM3) which are either explicitly annotated as negative or not present in the ground truth. We balance the training data by keeping the number of negative examples the same as the number of positive examples. These examples are then divided into 5-folds for cross-validation. We create these folds at the query level.

Baselines. We compare our proposed re-ranking approach `DREQ` to the following supervised state-of-the-art neural re-rankers: (1) **RoBERTa [37]**, (2) **DeBERTa [24]**, (3) **ELECTRA [7]**, (4) **ConvBERT [28]**, (5) **RankT5 [76]**, (6) **KNRM [63]**, (7) **ERNIE [72]**, (8) **EDRM [38]**. Furthermore, we also include an unsupervised entity-based baseline (9) **MaxSimCos** which scores documents using the maximum cosine similarity between every pair of query and document entity embedding. On TREC Robust 2004, we include the following additional (full-retrieval) baselines: (1) CEDR [39], (2) EQFE [11], (3) BERT-MaxP[8], and (4) PARADE [35]. All baselines are fine-tuned on the target datasets via 5-fold cross-validation using the binary cross-entropy loss.

4.3   Implementation Details

We use the `bert-base-uncased` model from HuggingFace to obtain the initial query and document/entity representations and fine-tune our model using the `CrossEntropyLoss` function from PyTorch. We use the Adam [32] optimizer with

Table 1: Overall results on TREC Robust 2004. Best results in bold. ▲ denotes significant improvement and ▼ denotes significant deterioration compared to ⋆. Paired-t-test at $p < 0.05$. **nDCG and Precision measures at cut-off rank 20**. Unavailable results denoted by "–". Baselines denotes by † are re-ranking the BM25+RM3 candidate set at the top.

		TREC Robust 2004 (Title)			TREC Robust 2004 (Desc)		
		MAP	nDCG	Prec	MAP	nDCG	Prec
	BM25+RM3	0.29⋆	0.44⋆	0.38⋆	0.28⋆	0.42⋆	0.37⋆
Non-entity	RankT5†	0.30	0.50▲	0.43▲	0.33▲	0.54▲	0.46▲
	RoBERTa†	0.29	0.47▲	0.41▲	0.33▲	0.54▲	0.46▲
	DeBERTa†	0.29	0.49▲	0.42▲	0.34▲	0.55▲	0.47▲
	ELECTRA†	0.27▼	0.45	0.39	0.29	0.49▲	0.41▲
	ConvBERT†	0.32▲	0.52▲	0.45▲	0.35▲	0.57▲	0.48▲
	KNRM†	0.11▼	0.18▼	0.16▼	0.08▼	0.14▼	0.12▼
	CEDR	0.37▲	0.55▲	0.48▲	0.40▲	0.60▲	0.52▲
	PARADE	0.30	0.53▲	0.45▲	0.30▲	0.56▲	0.47▲
	BERT-MaxP	0.32▲	0.48▲	0.42▲	0.31▲	0.49▲	0.22▼
Entity-based	ERNIE†	0.29	0.48▲	0.41▲	0.33▲	0.54▲	0.45▲
	EDRM†	0.07▼	0.10▼	0.09▼	0.05▼	0.07▼	0.07▼
	EQFE	0.33▼	0.42▼	0.38▼	–	–	–
	MaxSimCos†	0.17▼	0.26▼	0.24▼	0.13▼	0.20▼	0.18▼
	TREC Best	0.33▲	–	–	0.33▲	–	–
	DREQ	**0.57▲**	**0.75▲**	**0.73▲**	**0.55▲**	**0.78▲**	**0.71▲**

Table 2: Overall results on CODEC, TREC News 2021, and TREC Core 2018. **nDCG and Precision measures at cut-off rank 20**.

	CODEC			TREC News 2021			TREC Core 2018		
	MAP	nDCG	Prec	MAP	nDCG	Prec	MAP	nDCG	Prec
BM25+RM3	0.36⋆	0.38⋆	0.43⋆	0.47⋆	0.48⋆	0.58⋆	0.31⋆	0.45⋆	0.46⋆
RankT5	0.38▲	0.44▲	0.45▲	0.27▼	0.31▼	0.36▼	0.22▼	0.33▼	0.33▼
RoBERTa	0.36	0.38	0.43	0.47	0.52▲	0.58	0.26▼	0.36▼	0.39▼
DeBERTa	0.39▲	0.45▲	0.46▲	0.43▼	0.47	0.56	0.35▲	0.52▲	0.51▲
ELECTRA	0.32▼	0.32▼	0.41▼	0.23▼	0.46▼	0.53▼	0.24▼	0.35▼	0.36▼
ConvBERT	0.38▲	0.44▼	0.46▲	0.44▼	0.47	0.55	0.32	0.50▲	0.49▲
KNRM	0.30▼	0.30▼	0.34▼	0.14▼	0.15▼	0.19▼	0.11▼	0.14▼	0.16▼
ERNIE	0.39▲	0.46▲	0.47▲	0.48	0.53▼	0.60▲	0.34▲	0.52▲	0.51▲
EDRM	0.30▼	0.29▼	0.35▼	0.09▼	0.10▼	0.13▼	0.09▼	0.10▼	0.13▼
MaxSimCos	0.33▼	0.33▼	0.40▼	0.20▼	0.22▼	0.30▼	0.10▼	0.12▼	0.14▼
TREC Best	–	–	–	0.43▼	–	–	0.43▲	–	0.61▲
DREQ	**0.58▲**	**0.50▲**	**0.68▲**	**0.76▲**	**0.56▲**	**0.84▲**	**0.58▲**	**0.66▲**	**0.71▲**

a learning rate of 10^{-5} and batch size of 20. BERT layers were not frozen during fine-tuning. For document segmentation, we apply a 10-sentence sliding window with a 5-sentence stride using spaCy (v6.3.1). Embeddings (document text, and entity) and entity links are are cached off-line for lookup during inference.

5 Results and Discussions

5.1 Effectiveness of DREQ

In this section, we explore the following research question: **(RQ1)** *How does DREQ perform in comparison to state-of-the-art document re-ranking methods? Which type of queries does it help the most?* From Tables 1 and 2, we observe that DREQ outperforms all baselines in terms of all evaluation metrics on all datasets.
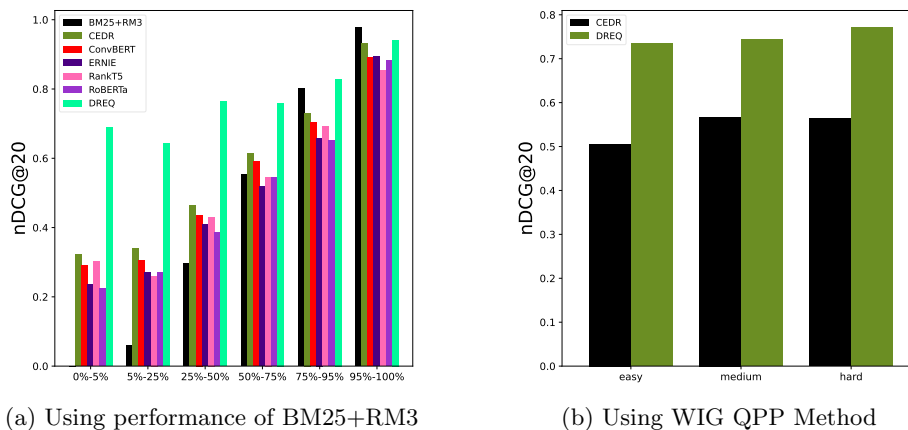
(a) Using performance of BM25+RM3

(b) Using WIG QPP Method

Fig. 2: Difficulty test for nDCG@20 on Robust04 (title). `DREQ` improves performance for the most difficult queries.

For example, on Robust04 (title), the best performing baseline is CEDR which obtains nDCG@20 = 0.55; however, `DREQ` improves performance by 36% over CEDR and achieves nDCG@20 = 0.75. While CEDR improves nDCG@20 of the candidate set (BM25+RM3) by 25% (from 0.44 to 0.55), our model `DREQ` improves this by 70% (from 0.44 to 0.75).

Query-level Analysis. We further delve into the source of performance improvements by analyzing the results at the query-level. We categorize the queries into different levels of difficulty using (a) the performance of the BM25+RM3 candidate ranking method, and (b) Weighted Information Gain (WIG) [75] Query Performance Prediction (QPP) method. Results shown in Figure 2.

In Figure 2a, we place the 5% most difficult queries (according to nDCG@20) for BM25+RM3, the candidate ranking method, on the left, and the 5% easiest ones on the right. Remarkably, `DREQ` enhances the performance of the most challenging queries (bins 0–25%). For example, there are no relevant documents among the top-20 for queries in bin 0–5% (nDCG@20=0.0); however, `DREQ` improves (helps) the performance of these queries, achieving an nDCG@20 of 0.70 (twice of CEDR which achieves 0.35). Specifically, we find that `DREQ` improves the performance of 210 of the 250 queries whereas CEDR only helps 160 queries.

In Figure 2b, we specifically study how the reranking pipeline `BM25+RM3 >> DREQ` compares to the best performing baseline pipeline `BM25+RM3 >> CEDR`. For this, we use a well-known QPP method called WIG which provides a measure of the effectiveness of a query in retrieving relevant documents: A higher WIG score means the query is more effective (easier) while a lower score means the query is less effective (harder). Based on this, we divide the query set into three levels of difficulty: easy, medium, and hard. Once again, we observe that compared to CEDR, `DREQ` improves performance for the most challenging queries.

Example. We further examine some of the challenging queries helped by `DREQ` from bin 0–5% in Figure 2a. One such query is "Behavioral Genetics".

For this query, a highly relevant document (according to NIST assessments) is placed at rank 434 in the candidate set. However, we find that `DREQ` promotes this document to rank 1 (CEDR places this document at rank 212). Upon analyzing the entity ranking for this query, we find that entities directly relevant to the query, such as *Sensorineural Hearing Loss*, are assigned higher scores, while non-relevant entities, such as *Home for the Holidays*, receive lower scores. By utilizing these entity scores to weigh the entity embeddings, the model can emphasize the most important entities for the query, thereby enhancing the contribution of these entities to the hybrid document embedding. This, in turn, helps the model understand the nuances of relevance for the query, thereby improving the document's ranking position. We discuss contribution of entities in Section 5.2.

Take Away. To answer **RQ1**, `DREQ` achieves new state-of-the-art results for the document re-ranking task on four document ranking benchmarks across five diverse query sets by promoting relevant documents to the top of the ranking, thereby improving the precision at the top ranks of the candidate set. Our experiments show that `DREQ` is particularly adept at handling the most challenging queries, a key advantage over competing models. This demonstrates the robustness of `DREQ`, as it not only improves the performance of "easy" queries but also significantly boosts the performance of the most "difficult" ones. This is crucial for real-world applications where queries may be of varying difficulty.

5.2   Contribution of Entities in `DREQ`

Entities lie at the heart of our approach. Hence, in this section, we ask: **(RQ2)** *What is the contribution of entities in `DREQ`? How does the performance change if we change the underlying entity ranking system?* Furthermore, **(RQ3)** *How does changing the entity weighing method alter the performance of `DREQ`?* In the following, **we discuss results with respect to only Robust04 (title)**.

Ablation Study. To evaluate the contribution of the entity component to our model's performance, we first conduct an ablation study by removing the entity embeddings and retraining the model using only the other components. Our results reveal that, when entities are removed from `DREQ`, there's a marked performance drop. The nDCG@20 score plunges to 0.26, a decrease from the 0.75 score of the full model, and even below the 0.44 of the BM25+RM3 candidate set. For a clearer context, it is worth noting that the highest nDCG@20 score achieved by the best baseline model, CEDR, is 0.55. Performing the difficulty test described above using the BM25+RM3 candidate ranking, we find that `DREQ` without entities helps 74 queries; in contrast, the full `DREQ` model helps 210.

These results underline the pivotal role of entities in `DREQ`. The marked reduction in performance, when compared to both the full model and the best baseline, indicates that entities are integral to the model's effectiveness in document ranking and relevance interpretation. The comparison with CEDR, the best performing baseline, underscores the significance of incorporating entities, as even CEDR fails to match the nDCG@20 score of the full `DREQ` model. This suggests that entities are not only essential for `DREQ` but could also be a valuable addition to other document ranking models to enhance their performance.

Alternative Entity Ranking. As we use the rank scores of entities to weigh the entity embeddings when learning the hybrid document embedding, we study the effect of the entity ranking system on the performance of `DREQ`. We replace our supervised BERT-based entity ranking system with (1) BM25, a sparse model and, (2) GEEER [19], a recently proposed, state-of-the-art, entity re-ranking method. GEEER first computes the (Wikipedia2Vec [69]) embedding-based score for an entity E in a given candidate set of entities as a weighted sum: $S_{emb}(E, Q) = \sum_{e \in Q} C(e) \cdot \cos(\mathbf{E}, \mathbf{e})$, where $C(e)$ is the confidence score of entity $e \in Q$ obtained from an entity linker. The final score is obtained by interpolating this embedding-based score with the score of the entity E obtained using a retrieval model. In this work, we use BM25 as the retrieval model in GEEER.

In the results, we find that incorporating the BM25 entity ranking into `DREQ` leads to a decline in performance, with the nDCG@20 dropping from 0.75 to 0.44. Similarly, substituting our BERT-based entity ranking system with GEEER causes a decrease in the nDCG@20 for `DREQ`, from 0.75 to 0.42. Our results underscore the importance of carefully selecting an entity ranking system that is well-suited to the specific requirements of the task. It is particularly important to note that merely using an off-the-shelf entity ranking system is insufficient; it is necessary to train the system specifically to predict entities that are likely to be mentioned in documents relevant to the query.

Alternative Entity Weighing. The `DREQ` model intrinsically uses entity scores, expressed as probabilities, to weight the entity embeddings within documents. Acknowledging the pivotal role these probabilistic weightings have in shaping the model's performance, we perform a series of experiments to examine various alternative weighting schemes.

First, **we remove entity weightings**, treating all entities with equivalent importance within the document's representation ($w_e = 1$ in Equation 1). This leads to a significant performance decrease in `DREQ`, from an nDCG@20 of 0.75 to 0.70. The difficulty test, using the BM25+RM3 candidate ranking, shows that `DREQ` using uniform weights helps 187 queries, whereas the original model helps 210. This underscores the pivotal role of entities in `DREQ`. Their embeddings, when appropriately weighted by relevance probabilities, significantly amplify the semantic alignment between queries and documents, attesting to the strategic decision to utilize them. Their importance becomes evident when the individual significance is disregarded, leading to a diluted representation that hampers the model's performance. Furthermore, it's clear that the ranking system's ability to discern and prioritize relevant entities is instrumental to `DREQ`'s success.

Additionally, we explore the **reciprocal rank (RR) of entities** as an alternative weighting mechanism for entity embeddings ($w_e = \frac{1}{rank(e)}$ in Equation 1). The results are telling: nDCG@20 plummets from 0.75 to 0.54. The RR method assumes a precipitous decrease in the importance of entities based on rank. However, this isn't always fitting. Take, for instance, the query "Black Bear Attacks". We find that the entity *George Parks Highway*, with a score of 0.99, is ranked 71st, while *Cantwell, Alaska*, with a nearly identical score of 0.98, stands at 106th rank. Such negligible probability differences juxtaposed against

substantial rank disparities highlight the potential limitations of the RR system. It underscores the risk of undervaluing entities that, by probability, are nearly as pertinent as top-tier entities.

Take Away. To answer **RQ2**, the significant role of entities in the performance of `DREQ` is undeniable, as evident from the ablation study and alternative entity ranking experiments. The experiments indicate that entities are integral to the model's effectiveness, and that the entity ranking system's ability to discern and prioritize query-relevant entities is instrumental to the success of `DREQ`. However, it is essential to carefully select and train the entity ranking system to predict entities likely to be mentioned in documents relevant to the query.

In addressing **RQ3**, our experiments unveil crucial insights. The inherent design of `DREQ`, which uses entity scores as probabilities to weight the entity embeddings, stands out as a potent approach, demonstrated by superior performance metrics. When entities are uniformly weighted, stripping them of their probabilistic distinctions, the model's performance takes a tangible hit, underscoring the model's dependence on the nuanced gradation of entity relevance. The experiment with RR as weights further cements this observation. While RR attempts to rank entities based on their ordinal positions, it fails to adequately capture the subtleties in actual relevance, particularly when entities with near-identical probabilities have significantly differing ranks. The experiments not only re-emphasize the instrumental role of entities in `DREQ` (**RQ2**), but also the imperativeness of a precise entity ranking and weighting mechanism.

6   Conclusion

We introduce `DREQ`, an entity-oriented document re-ranking model that taps into the semantic richness of entities in candidate documents. While many neural IR models overlook the differential importance of entities within documents, `DREQ` prioritizes query-relevant entities, crafting a nuanced, query-specific entity-centric representation. This is then melded with the document's text-centric representation to form a *hybrid* embedding, capturing *query-specific insights*. We provide compelling evidence of the effectiveness of `DREQ` in the document re-ranking task across multiple benchmarks and diverse query sets. `DREQ` achieves new state-of-the-art results by enhancing precision at the top ranks of the candidate set. Its proficiency in handling challenging queries demonstrates its robustness and underscores its potential for real-world applications where queries may vary in difficulty. We also demonstrate the pivotal role of entities, and the importance of meticulously selecting and weighing entities in `DREQ`'s performance.

Overall, our work contributes significantly to the ongoing efforts to enhance the effectiveness of information retrieval systems. The `DREQ` model, with its innovative approach to incorporating entities and its proven ability to outperform existing methods, presents a promising avenue for further exploration and development in the field of information retrieval.

References

1. Akkalyoncu Yilmaz, Z., Yang, W., Zhang, H., Lin, J.: Cross-domain modeling of sentence-level evidence for document retrieval. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3490–3496. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1352, https://aclanthology.org/D19-1352

2. Balog, K., Bron, M., De Rijke, M.: Query modeling for entity search based on terms, categories, and examples. ACM Transactions on Information Systems **29**(4) (Dec 2011). https://doi.org/10.1145/2037661.2037667, https://doi.org/10.1145/2037661.2037667

3. Cao, N.D., Izacard, G., Riedel, S., Petroni, F.: Autoregressive entity retrieval. In: International Conference on Learning Representations (2021), https://openreview.net/pdf?id=5k8F6UU39V

4. Chatterjee, S., Dietz, L.: Entity Retrieval Using Fine-Grained Entity Aspects. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1662–1666. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463035, https://doi.org/10.1145/3404835.3463035

5. Chatterjee, S., Dietz, L.: Bert-er: Query-specific bert entity representations for entity ranking. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1466–1477. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3531944, https://doi.org/10.1145/3477495.3531944

6. Ciglan, M., Nørvåg, K., Hluchý, L.: The semsets model for ad-hoc semantic list search. In: Proceedings of the 21st International Conference on World Wide Web. p. 131–140. WWW '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2187836.2187855, https://doi.org/10.1145/2187836.2187855

7. Clark, K., Luong, M., Le, Q.V., Manning, C.D.: ELECTRA: pre-training text encoders as discriminators rather than generators. CoRR **abs/2003.10555** (2020), https://arxiv.org/abs/2003.10555

8. Dai, Z., Callan, J.: Deeper text understanding for IR with contextual neural language modeling. CoRR **abs/1905.09217** (2019), http://arxiv.org/abs/1905.09217

9. Dai, Z., Callan, J.: Context-aware term weighting for first stage passage retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1533–1536. SIGIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401204, https://doi.org/10.1145/3397271.3401204

10. Dai, Z., Xiong, C., Callan, J., Liu, Z.: Convolutional neural networks for soft-matching n-grams in ad-hoc search. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. p. 126–134. WSDM '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3159652.3159659, https://doi.org/10.1145/3159652.3159659

11. Dalton, J., Dietz, L., Allan, J.: Entity query feature expansion using knowledge base links. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. pp. 365–374. ACM (2014)
12. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
13. Dietz, L.: ENT Rank: Retrieving Entities for Topical Information Needs through Entity-Neighbor-Text Relations. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 215–224. SIGIR'19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3331184.3331257, https://doi.org/10.1145/3331184.3331257
14. Dietz, L., Gamari, B., Dalton, J., Craswell, N.: Trec complex answer retrieval overview. In: TREC (2018)
15. Dietz, L., Verma, M., Radlinski, F., Craswell, N.: Trec complex answer retrieval overview. In: Proceedings of Text REtrieval Conference (TREC) (2017)
16. Ensan, F., Bagheri, E.: Document retrieval model through semantic linking. In: Proceedings of the 10th ACM International Conference on Web Search and Data Mining. p. 181–190. WSDM '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3018661.3018692, https://doi.org/10.1145/3018661.3018692
17. Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research **34**, 443–498 (2009)
18. Garigliotti, D., Balog, K.: On type-aware entity retrieval. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. p. 27–34. ICTIR '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3121050.3121054, https://doi.org/10.1145/3121050.3121054
19. Gerritse, E.J., Hasibi, F., de Vries, A.P.: Graph-Embedding Empowered Entity Retrieval. In: Advances in Information Retrieval, Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020). pp. 97–110. Lecture Notes in Computer Science, Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_7, https://doi.org/10.1007/978-3-030-45439-5_7
20. Gerritse, E.J., Hasibi, F., de Vries, A.P.: Entity-aware transformers for entity search. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1455–1465. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3531971, https://doi.org/10.1145/3477495.3531971
21. Graus, D., Tsagkias, M., Weerkamp, W., Meij, E., de Rijke, M.: Dynamic collective entity representations for entity ranking. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. p. 595–604. WSDM '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2835776.2835819, https://doi.org/10.1145/2835776.2835819
22. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. p. 55–64. CIKM '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2983323.2983769, https://doi.org/10.1145/2983323.2983769

23. Hasibi, F., Balog, K., Bratsberg, S.E.: Exploiting entity linking in queries for entity retrieval. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. p. 209–218. ICTIR '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2970398.2970406, https://doi.org/10.1145/2970398.2970406

24. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced BERT with disentangled attention. CoRR **abs/2006.03654** (2020), https://arxiv.org/abs/2006.03654

25. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. p. 2333–2338. CIKM '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2505515.2505665, https://doi.org/10.1145/2505515.2505665

26. Hui, K., Yates, A., Berberich, K., de Melo, G.: PACRR: A position-aware neural IR model for relevance matching. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1049–1058. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1110, https://aclanthology.org/D17-1110

27. Hui, K., Yates, A., Berberich, K., de Melo, G.: Co-pacrr: A context-aware neural ir model for ad-hoc retrieval. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. p. 279–287. WSDM '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3159652.3159689, https://doi.org/10.1145/3159652.3159689

28. Jiang, Z., Yu, W., Zhou, D., Chen, Y., Feng, J., Yan, S.: Convbert: Improving BERT with span-based dynamic convolution. CoRR **abs/2008.02496** (2020), https://arxiv.org/abs/2008.02496

29. Kaptein, R., Serdyukov, P., De Vries, A., Kamps, J.: Entity ranking using wikipedia as a pivot. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. p. 69–78. CIKM '10, Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1871437.1871451, https://doi.org/10.1145/1871437.1871451

30. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.550, https://aclanthology.org/2020.emnlp-main.550

31. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 39–48. SIGIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401075, https://doi.org/10.1145/3397271.3401075

32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

33. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia–a large-scale,

multilingual knowledge base extracted from wikipedia. Semantic Web **6**, 167–195 (2015). https://doi.org/10.3233/SW-140134

34. Li, C., Sun, Y., He, B., Wang, L., Hui, K., Yates, A., Sun, L., Xu, J.: NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4482–4491. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1478, https://aclanthology.org/D18-1478

35. Li, C., Yates, A., MacAvaney, S., He, B., Sun, Y.: PARADE: passage representation aggregation for document reranking. CoRR **abs/2008.09093** (2020), https://arxiv.org/abs/2008.09093

36. Liu, X., Fang, H.: Latent entity space: a novel retrieval approach for entity-bearing queries. Information Retrieval Journal **18**(6), 473–503 (2015)

37. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692

38. Liu, Z., Xiong, C., Sun, M., Liu, Z.: Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2395–2405. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1223, https://aclanthology.org/P18-1223

39. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: CEDR: contextualized embeddings for document ranking. CoRR **abs/1904.07094** (2019), http://arxiv.org/abs/1904.07094

40. Mackie, I., Owoicho, P., Gemmell, C., Fischer, S., MacAvaney, S., Dalton, J.: Codec: Complex document and entity collection. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3067–3077. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3531712, https://doi.org/10.1145/3477495.3531712

41. Manotumruksa, J., Dalton, J., Meij, E., Yilmaz, E.: Crossbert: A triplet neural architecture for ranking entity properties. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2049–2052. SIGIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401265, https://doi.org/10.1145/3397271.3401265

42. Metzler, D., Croft, W.B.: A markov random field model for term dependencies. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 472–479. SIGIR '05, Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1076034.1076115, https://doi.org/10.1145/1076034.1076115

43. Mitra, B., Craswell, N.: An updated duet model for passage re-ranking. CoRR **abs/1903.07666** (2019), http://arxiv.org/abs/1903.07666

44. Nalisnick, E., Mitra, B., Craswell, N., Caruana, R.: Improving document ranking with dual word embeddings. In: Proceedings of the 25th International Conference Companion on World Wide Web. p. 83–84. WWW '16 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2016). https://doi.org/10.1145/2872518.2889361, https://doi.org/10.1145/2872518.2889361

45. Naseri, S., Dalton, J., Yates, A., Allan, J.: Ceqe: Contextualized embeddings for query expansion. In: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part I. p. 467–482. Springer-Verlag, Berlin, Heidelberg (2021). https://doi.org/10.1007/978-3-030-72113-8_31, https://doi.org/10.1007/978-3-030-72113-8_31

46. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. CoRR **abs/1611.09268** (2016), http://arxiv.org/abs/1611.09268

47. Nikolaev, F., Kotov, A.: Joint word and entity embeddings for entity retrieval from a knowledge graph. In: Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I. p. 141–155. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-45439-5_10, https://doi.org/10.1007/978-3-030-45439-5_10

48. Nikolaev, F., Kotov, A., Zhiltsov, N.: Parameterized fielded term dependence models for ad-hoc entity retrieval from knowledge graph. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 435–444. SIGIR '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2911451.2911545, https://doi.org/10.1145/2911451.2911545

49. Nogueira, R.F., Cho, K.: Passage re-ranking with BERT. CoRR **abs/1901.04085** (2019), http://arxiv.org/abs/1901.04085

50. Nogueira, R.F., Yang, W., Cho, K., Lin, J.: Multi-stage document ranking with BERT. CoRR **abs/1910.14424** (2019), http://arxiv.org/abs/1910.14424

51. Nogueira, R.F., Yang, W., Lin, J., Cho, K.: Document expansion by query prediction. CoRR **abs/1904.08375** (2019), http://arxiv.org/abs/1904.08375

52. Piccinno, F., Ferragina, P.: From tagme to wat: A new entity annotator. In: Proceedings of the First International Workshop on Entity Recognition & Disambiguation. p. 55–62. ERD '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2633211.2634350, https://doi.org/10.1145/2633211.2634350

53. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR **abs/1910.10683** (2019), http://arxiv.org/abs/1910.10683

54. Raviv, H., Carmel, D., Kurland, O.: A ranking framework for entity oriented search using markov random fields. In: Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search. JIWES '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2379307.2379308, https://doi.org/10.1145/2379307.2379308

55. Raviv, H., Kurland, O., Carmel, D.: Document retrieval using entity-based language models. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 65–74. SIGIR '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2911451.2911508, https://doi.org/10.1145/2911451.2911508

56. Schuhmacher, M., Dietz, L., Paolo Ponzetto, S.: Ranking Entities for Web Queries Through Text and Knowledge. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. p.

1461–1470. CIKM '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806480, https://doi.org/10.1145/2806416.2806480

57. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2), 443–460 (2015). https://doi.org/10.1109/TKDE.2014.2327028

58. Tonon, A., Demartini, G., Cudré-Mauroux, P.: Combining inverted indices and structured search for ad-hoc object retrieval. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 125–134. SIGIR '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2348283.2348304, https://doi.org/10.1145/2348283.2348304

59. Tran, H.D., Yates, A.: Dense retrieval with entity views. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. p. 1955–1964. CIKM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3511808.3557285, https://doi.org/10.1145/3511808.3557285

60. Wang, X., MacDonald, C., Tonellotto, N., Ounis, I.: Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. ACM Trans. Web **17**(1) (jan 2023). https://doi.org/10.1145/3572405, https://doi.org/10.1145/3572405

61. Xiong, C., Callan, J.: Esdrank: Connecting query and documents through external semi-structured data. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management. pp. 951–960. CIKM '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806456, http://doi.acm.org/10.1145/2806416.2806456

62. Xiong, C., Callan, J., Liu, T.Y.: Word-entity duet representations for document ranking. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 763–772. SIGIR '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3077136.3080768, https://doi.org/10.1145/3077136.3080768

63. Xiong, C., Dai, Z., Callan, J., Liu, Z., Power, R.: End-to-end neural ad-hoc ranking with kernel pooling. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 55–64. SIGIR '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3077136.3080809, https://doi.org/10.1145/3077136.3080809

64. Xiong, C., Liu, Z., Callan, J., Hovy, E.: Jointsem: Combining query entity linking and entity based document ranking. In: Proceedings of the 2017 ACM SIGIR Conference on Information and Knowledge Management. p. 2391–2394. CIKM '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3132847.3133048, https://doi.org/10.1145/3132847.3133048

65. Xiong, C., Liu, Z., Callan, J., Liu, T.Y.: Towards better text understanding and retrieval through kernel entity salience modeling. In: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 575–584. SIGIR '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3209978.3209982, https://doi.org/10.1145/3209978.3209982

66. Xiong, C., Power, R., Callan, J.: Explicit semantic ranking for academic search via knowledge graph embedding. In: Proceedings of the 26th International Conference on World Wide Web. p. 1271–1279. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). https://doi.org/10.1145/3038912.3052558, https://doi.org/10.1145/3038912.3052558

67. Xiong, C., Power, R., Callan, J.: Explicit semantic ranking for academic search via knowledge graph embedding. In: Proceedings of the 26th International Conference on World Wide Web. p. 1271–1279. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). https://doi.org/10.1145/3038912.3052558, https://doi.org/10.1145/3038912.3052558

68. Xiong, L., Xiong, C., Li, Y., Tang, K., Liu, J., Bennett, P.N., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. CoRR **abs/2007.00808** (2020), https://arxiv.org/abs/2007.00808

69. Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., Matsumoto, Y.: Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 23–30. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.4, https://aclanthology.org/2020.emnlp-demos.4

70. Yamada, I., Shindo, H., Takefuji, Y.: Representation learning of entities and documents from knowledge base descriptions. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 190–201. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://aclanthology.org/C18-1016

71. Yu, H., Xiong, C., Callan, J.: Improving query representations for dense retrieval with pseudo relevance feedback. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. p. 3592–3596. CIKM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3459637.3482124, https://doi.org/10.1145/3459637.3482124

72. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: Enhanced Language Representation with Informative Entities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1441–1451. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1139, https://aclanthology.org/P19-1139

73. Zheng, Z., Hui, K., He, B., Han, X., Sun, L., Yates, A.: BERT-QE: Contextualized Query Expansion for Document Re-ranking. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4718–4728. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.424, https://aclanthology.org/2020.findings-emnlp.424

74. Zhiltsov, N., Kotov, A., Nikolaev, F.: Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 253–262. SIGIR '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2766462.2767756, https://doi.org/10.1145/2766462.2767756

75. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 543–550. SIGIR '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1277741.1277835, https://doi.org/10.1145/1277741.1277835

76. Zhuang, H., Qin, Z., Jagerman, R., Hui, K., Ma, J., Lu, J., Ni, J., Wang, X., Bendersky, M.: Rankt5: Fine-tuning t5 for text ranking with ranking losses. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2308–2313. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3592047, https://doi.org/10.1145/3539618.3592047