Information Processing and Management 59 (2022) 103026

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

INFORMATION PROCESSING & MANAGEMENT

An International Journal

Improving zero-shot retrieval using dense external expansion

Xiao Wang, Craig Macdonald *, Iadh Ounis

University of Glasgow, School of Computing Science, Lilybank Gardens, Glasgow, G12 8QQ, United Kingdom

ARTICLE INFO

Keywords:
Query expansion
Pseudo-relevance feedback
Dense retrieval
Information retrieval

ABSTRACT

Pseudo-relevance feedback (PRF) is a classical technique to improve search engine retrieval effectiveness, by closing the vocabulary gap between users' query formulations and the relevant documents. While PRF is typically applied on the same *target corpus* as the final retrieval, in the past, external expansion techniques have sometimes been applied to obtain a high-quality pseudo-relevant feedback set using the *external corpus*. However, such external expansion approaches have only been studied for sparse (BoW) retrieval methods, and its effectiveness for recent dense retrieval methods remains under-investigated. Indeed, dense retrieval approaches such as ANCE and ColBERT, which conduct similarity search based on encoded contextualised query and document embeddings, are of increasing importance. Moreover, pseudo-relevance feedback mechanisms have been proposed to further enhance dense retrieval effectiveness. In particular, in this work, we examine the application of dense external expansion to improve *zero-shot* retrieval effectiveness, i.e. evaluation on corpora without further training. Zero-shot retrieval experiments with six datasets, including two TREC datasets and four BEIR datasets, when applying the MSMARCO passage collection as external corpus, indicate that obtaining external feedback documents using ColBERT can significantly improve NDCG@10 for the sparse retrieval (by upto 28%) and the dense retrieval (by upto 12%). In addition, using ANCE on the external corpus brings upto 30% NDCG@10 improvements for the sparse retrieval and upto 29% for the dense retrieval.

1. Introduction

Pseudo-Relevance Feedback (PRF) is a well known technique in information retrieval systems that has demonstrated that the relevance signal from the returned top documents is useful for improving the retrieval effectiveness for the prevalent bag-of-words (BoW, or *sparse*) retrieval models, such as BM25 (Azad & Deepak, 2019). By expanding the original query with additional information extracted from the returned documents, the vocabulary mismatch problem can be alleviated (Abdul-Jaleel et al., 2004; Amati & Van Rijsbergen, 2002). However, if the used vocabulary in the pseudo-relevance feedback documents is limited, the expansion terms may not identify more relevant documents. Furthermore, for some hard queries, the feedback set may contain non-relevant documents, thus causing topic drift from incorrect expansion terms.

Instead, it has previously been shown that high-quality external corpora can also produce feedback documents with complementary information (Diaz & Metzler, 2006; Kwok & Chan, 1998; Peng, He and Ounis, 2009; Peng, Macdonald, He and Ounis, 2009; Xu, Jones, & Wang, 2009).[1] Thus, different query related words can be extracted to reformulate, typically expand, the original query. Usually, the additional collection is referred to as the *external collection* while the local collection used is referred to as the

* Corresponding author.

E-mail addresses: x.wang.8@research.gla.ac.uk (X. Wang), craig.macdonald@glasgow.ac.uk (C. Macdonald), iadh.ounis@glasgow.ac.uk (I. Ounis).

[1] This technique is also known as *collection enrichment* (Kwok & Chan, 1998); we prefer the modern nomenclature of *external expansion* (Diaz & Metzler, 2006), which is more easily relatable to its definition.

https://doi.org/10.1016/j.ipm.2022.103026
Received 20 May 2022; Received in revised form 7 July 2022; Accepted 8 July 2022
Available online 2 August 2022

target collection. However, such *external expansion* approaches have only been studied for sparse (BoW) retrieval methods, and its effectiveness for recent dense retrieval methods remains uninvestigated.

Indeed, large pre-trained contextualised language models (LM), such as BERT (Devlin, Chang, Lee, & Toutanova, 2019), have brought marked benefits to retrieval effectiveness (Lin, Nogueira, & Yates, 2021). BERT-based rerankers allow the contextualised re-scoring of a set of candidate documents provided by a traditional sparse retrieval model such as BM25, or even BM25+PRF (Lin et al., 2021; Naseri, Dalton, Yates, & Allan, 2021; Wang, Macdonald, & Tonellotto, 2021). Using a contextualised embedding space can address both word-mismatch and polysemy, as different meanings will obtain different embeddings, while dissimilar words with similar meanings will have similar embeddings. Building on the use of BERT-like models as rerankers, dense retrieval approaches perform end-to-end retrieval, relying on the contextualised query and document embedding(s), without the need for a classical inverted index and a BoW model. Two families of dense retrieval models have emerged (Macdonald, Tonellotto, & Ounis, 2021): in *single representation* dense retrieval approaches, such DPR (Karpukhin et al., 2020) and ANCE (Xiong et al., 2021), each query and document is regarded as a whole when encoding into a single contextualised embedding. In contrast, in a *multiple representation* dense retrieval – exemplified by ColBERT (Khattab & Zaharia, 2020) – contextualised embeddings are produced for each token of the queries or documents.

As the training of such dense retrieval models require large amounts of training data, there may not be enough training data for smaller corpora to train effective domain-specific dense retrieval models. An attractive solution is to *transfer* sufficiently trained dense retrieval models from large corpora to smaller domains, which is called *zero-shot*. For instance, MacAvaney, Yates, Cohan, and Goharian (2019) showed that an effective BERT reranker could be trained using MSMARCO passage ranking data, for adhoc search, as well as COVID-related literature (MacAvaney, Cohan and Goharian, 2020). In contrast, attempts using zero-shot dense retrieval models have under-performed compared to existing models (Chen, Zhang, Lu, Bendersky and Najork, 2022; Thakur, Reimers, Rücklé, Srivastava, & Gurevych, 2021). To mitigate this domain shift between the external trained collection and the target collection, we propose to employ the external expansion from the external high quality collection thus improving zero-shot retrieval performance.

On the one hand, we investigate the benefit to a classical BoW PRF model that is supported by a pseudo-relevant feedback set obtained from different dense retrieval approaches. Moreover, we also study the effectiveness of external dense expansion on the dense retrieval. This facilitates an investigation into how external expansion changes the semantic manner of retrieval. Finally, we investigate whether the sparse external retrieval can produce feedback information that is useful for the dense retrieval.

In summary, our work makes the following contributions: we investigate external expansion when mixing sparse & dense retrieval paradigms (including both single representation and multiple representation dense retrieval); sparse and zero-shot dense retrieval experiments are conducted on two classical TREC test collections (Robust04 & WT10G); We also conduct zero-shot evaluation on four BEIR datasets (Thakur et al., 2021) (namely DBPedia, NFCorpus, TREC-COVID and Touché-2020) - a set of datasets selected for evaluating zero-shot evaluation; We deploy two sparse weighting models (BM25 & DPH), two sparse PRF approaches (RM3 & Bo1), two dense retrieval models (ANCE & ColBERT), and two dense RPF approaches (ANCE-PRF & ColBERT-PRF); we analyse the propensity for a multiple representation PRF technique to perform semantic vs. exact token matching, under normal and external PRF conditions.

The main findings of this work are: (1) high quality feedback documents obtained using multiple representation dense retrieval, namely ColBERT, from a high-quality external collection can significantly improve sparse retrieval for both Robust04 (by 12% for NDCG@10) and WT10G (by 28% for NDCG@10); (2) significant sparse retrieval effectiveness improvements are also observed when performing expansion using external feedback documents obtained using a single representation dense retrieval, namely ANCE; (3) extracting PRF documents from an external collection using ColBERT or ANCE for dense retrieval can significantly outperform the zero-shot dense retrieval models on target, indicating the utility of the dense external expansion to improve the effectiveness of zero-shot dense retrieval.

The remainder of this paper is as follows: Section 2 discusses related works; Section 3 presents a framework for portraying retrieval and pseudo-relevance feedback techniques. Section 4 instantiates this framework for external pseudo-relevance expansion; Section 5 elicits our research questions; Section 6 and Section 7 present the experimental setup and results of this work. Finally, we provide the concluding remarks in Section 8.

2. Related work

Pseudo-relevance feedback is a classical and popular approach to alleviate the vocabulary mismatch problem (Abdul-Jaleel et al., 2004; Amati & Van Rijsbergen, 2002; Azad & Deepak, 2019; Croft, Metzler, & Strohman, 2010; Croft et al., 2010; He & Ounis, 2007; Lavrenko & Croft, 2001; Lioma & Ounis, 2008; Pan et al., 2022; Wang et al., 2020; Wong, Luk, Leong, Ho, & Lee, 2008). Many approaches have been proposed by identifying expansion terms from the top relevance documents, such as the Rocchio algorithm (Rocchio, 1971), the RM3 (Abdul-Jaleel et al., 2004) relevance language model, the DFR Bo1 query expansion model (Amati & Van Rijsbergen, 2002). and the recent CEQE (Naseri et al., 2021) model. Neural PRF models, such as NPRF (Li et al., 2018; Wang et al., 2020) and BERT-QE (Zheng et al., 2020) incorporate the PRF information for modelling the query document similarity, but only act as rerankers, and thus cannot enhance recall. Other recent efforts attempted to address the vocabulary mismatch problem using neural-based models were to augment the sparse document representation. For instance, both doc2query (Nogueira, Lin and Epistemic, 2019; Nogueira, Yang, Lin and Cho, 2019) and DeepImpact (Mallia, Khattab, Suel, & Tonellotto, 2021) use a sequence-to-sequence model to generate terms that are likely to appear in queries for the document. In addition, inspired by EPIC (MacAvaney et al., 2020), SPLADE (Formal, Piwowarski, & Clinchant, 2021) which performs document expansion by using the masked language model directly to predict alternative terms that could appear in the document, to create an augmented inverted index.

On the other hand, instead of only obtaining the pseudo-relevance feedback documents from the target corpora, another thread of related research focuses on bringing in more high quality feedback documents from an external corpora. For instance, Kwok and Chan (1998) proposed an external expansion approach, which borrowed similar documents from an external collection when the quality of the initial returned documents is poor, in order to improve PRF. External expansion techniques were popular in the top runs in the TREC Robust Track (Voorhees, 2005, 2006). Indeed, Diaz and Metzler (2006) studied the collection size and other characteristics that an external corpora should have to provide the useful external information for a target collection. Meanwhile, Peng, He et al. (2009) and Peng, Macdonald et al. (2009) further studied a selective external expansion approach based on the query performance prediction, where external expansion was only applied when the external corpus was predicted to better answer the query than the target corpus. However, these external corpora based query reformulation techniques are only deployed in the classical sparse retrieval paradigm and have not yet been explored for the newer dense retrieval paradigm.

Indeed, different from traditional sparse retrieval, which mainly relies on the statistical information to model the similarity between queries and documents, dense retrieval leverages high-dimensional embedded representations to measure the contextualised relevance scores between queries and documents. In particular, single representation approaches such as DPR (Karpukhin et al., 2020) and ANCE (Xiong et al., 2021) use the approximate nearest neighbour (ANN) approaches (such as FAISS Johnson, Douze, & Jégou, 2019) to identify document embeddings most similar to a given query embedding. In multiple representation dense retrieval, exemplified by ColBERT (Khattab & Zaharia, 2020), each token of queries and documents are encoded into embeddings; an ANN index is then used to identify the most relevant document embeddings for each of the query embeddings. Due to the compression of the ColBERT ANN index, ColBERT re-ranks retrieved documents to obtain an effective ranking.

More recently, initial investigations have taken place into how pseudo-relevance feedback information can improve the effectiveness of dense retrieval. For instance, ColBERT-PRF (Wang et al., 2021) identifies representative and important embeddings from the pseudo-relevant set to expand the original query embeddings. ColBERT-PRF was shown to be very effective, exhibiting performances as high as the top-ranked group in terms of MAP and NDCG@10 in the TREC 2019 Deep Learning track. In contrast, instead of performing query expansion, ANCE-PRF (Li et al., 2021; Yu, Xiong and Callan, 2021) employs the pseudo-relevance feedback information within a retrained query encoder to capture the feedback information while encoding the query into a single high-dimensional embedding. Both ColBERT-PRF and ANCE-PRF aim to capture the useful information from the pseudo-relevance feedback documents to obtain query representations of the two PRF models, however they differ in the exact query representation, by using multiple representation and single representation query embeddings, respectively. We summarise the overall landscape of pseudo-relevance feedback and other related approaches in the literature in Table 1. Thus in this work, we build on top of the ColBERT-PRF and ANCE-PRF models, respectively, to investigate the application of external expansion in PRF for dense retrieval. In addition, we also study external expansion in a hybrid framework of the dense (ANCE and ColBERT) and sparse retrieval approaches.

3. Ranking & pseudo-relevance feedback models

3.1. Definitions & sparse models

Given some query formulation $q \in Q$, let $Ret(\mathcal{I}, k)(q) \rightarrow R$ denotes a retrieval process that takes the query q as input, and returns a ranking of k documents.[2] R, obtained from index $\mathcal{I}$; Further, given some set of ranked documents R, let a pseudo-relevance feedback process be denoted as $PRF(\mathcal{I}, \theta)(R) \rightarrow q$, where θ are the parameters of the process. Typically, a PRF process examines the ranked documents in the index, and in turn produces a reformulated query. This definition is sufficiently general enough to encompass classical pseudo-relevance feedback processes (e.g. RM3 or Bo1) on sparse retrieval models (e.g. BM25 or DPH) and, as we will show, dense retrieval models and pseudo-relevance feedback settings. To ease notation, let $\overset{x}{\gg}$ denote passing the output of one process, with type x, as input to another, where $x = R$ indicates the type is a ranking of documents and $x = Q$ indicates the type is a query. In this way, a classical PRF process, which uses a pseudo-relevance feedback set of k' documents, can be obtained using:

$$Ret_{BM25}(\mathcal{I}, k') \overset{R}{\gg} PRF_{RM3}(\mathcal{I}, \theta) \overset{Q}{\gg} Ret_{BM25}(\mathcal{I}, k). \tag{1}$$

Next, we discuss dense retrieval, and PRF approaches for both single & multiple representation dense retrieval frameworks.

3.2. ColBERT end-to-end dense retrieval

ColBERT-PRF applies the pseudo-relevance feedback mechanism on the ColBERT multiple representation dense retrieval model. As highlighted earlier, in multiple representation dense retrieval, the queries and the documents are each represented as the contextualised embeddings for each token. Thus, a query q can be represented as a set of $|q|$ embeddings $\{\phi_{q_1}, \ldots, \phi_{q_{|q|}}\}$. Typically $|q| = 32$, with unused query embeddings being used for "query augmentation", which can increase the amount of matching between queries and documents (Khattab & Zaharia, 2020). Similarly, a document d can be represented as a set of $|d|$ embeddings $\{\phi_{d_1}, \ldots, \phi_{d_{|d|}}\}$. The document embeddings are pre-computed and can be encoded into a FAISS (Johnson et al., 2019) ANN index,

[2] To aid readability, we use 'document' and 'passage' interchangeably in our explanations of the models in Sections 3 and 4.

Table 1
Representative pseudo-relevance feedback and related approaches in the literature, organised into two dimensions: the task of the approach and the type of index it conducted on.

Task	Index	
	Sparse index	Dense index
Query expansion	Rocchio (1971) Bo1 Amati and Van Rijsbergen (2002) RM3 (Abdul-Jaleel et al., 2004) Sparse external expansion (Peng, He et al., 2009) CEQE (Naseri et al., 2021) PGT (Yu, Dai and Callan, 2021)	ColBERT-PRF (Wang et al., 2021) ANCE-PRF (Yu, Xiong et al., 2021)
Document expansion	Doc2Query (Nogueira, Lin et al., 2019) DocT5Query (Nogueira, Yang et al., 2019) DeepCT (Dai & Callan, 2020) EPIC (MacAvaney, Nardini et al., 2020) SPLADE (Formal et al., 2021) DeepImpact (Mallia et al., 2021)	
Reranking	NPRF (Li et al., 2018) BERT-QE (Zheng et al., 2020) Co-BERT (Chen et al., 2022)	Vector-PRF (Li, Mourad, Zhuang, Koopman and Zuccon, 2021)

which allows the approximate nearest neighbour search using the query embeddings over the encoded document embeddings. More specifically, the ColBERT (Khattab & Zaharia, 2020) model follows a two stage retrieval framework, where a list of candidate documents is retrieved using an approximate nearest neighbour search followed by a more computationally demanding exact scoring stage. During the approximate nearest neighbour search stage, for each query token embedding, ColBERT returns a set of similar document embeddings. Then after mapping the returned document embeddings back to their docid, a set of k candidate documents is obtained. During the exact scoring stage, the similarity between a query q and a document d is computed by summing the maximum similarity scores between its query token embeddings and document token embeddings, as follows:

$$s(q,d) = \sum_{i=1}^{|q|} \max_{j=1,\ldots,|d|} \sigma_i \phi_{q_i}^T \phi_{d_j}, \tag{2}$$

where σ_i is a scalar indicator of the importance of the query token, usually 1. Using the notation of Section 3.1, let $Ret_{ColBERT\,E2E}(\mathcal{I},k)$ be a short-hand for the end-to-end dense retrieval, i.e. an ANN stage followed by maximum similarity reranker of Eq. (2):

$$Ret_{ColBERT}(\mathcal{I},k) = Ret_{ANN\ ColBERT}(\mathcal{I},k) \overset{R}{\gg} MaxSim(\mathcal{I}). \tag{3}$$

3.3. ColBERT-PRF

Following a PRF setup, ColBERT-PRF (Wang et al., 2021) is applied on a set of k' pseudo-relevant feedback documents, as follows:

(a) After obtaining the document embeddings from the index for each of the k' feedback documents, KMeans clustering is applied upon the embeddings to obtain representative (centroid) embeddings. These centroid embeddings represent contextualised tokens that appear regularly in the feedback documents.

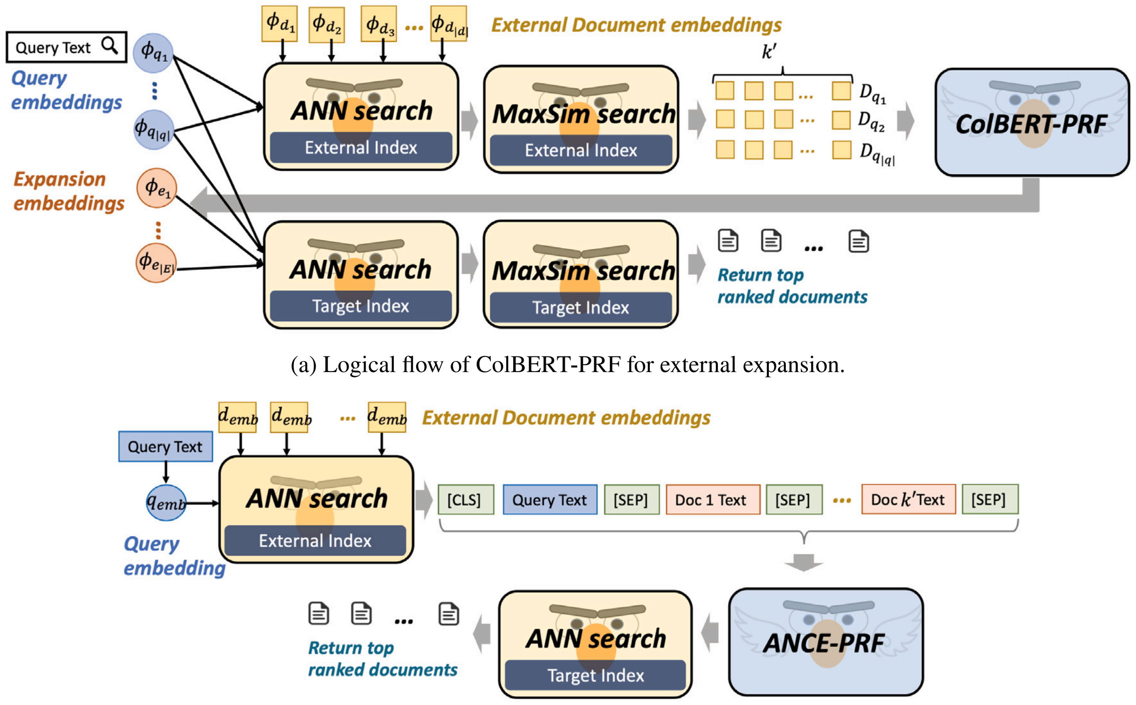
(b) To identify the most discriminative embeddings among the centroid embeddings, ColBERT-PRF employs FAISS to find the most likely token for a given representative embedding. Inverse Document Frequency (IDF) is then employed to measure the importance of each representative embedding based on the frequency of the corresponding token in the index; the most discriminative f_e embeddings are selected as the expansion embeddings ϕ_e.

(c) The final weight of an expansion embedding ϕ_i, σ_i – as used in Eq. (2) – encapsulates both the IDF and a hyperparameter β that controls the emphasis of the expansion embeddings in the final retrieval process in comparison to the original query embeddings, i.e. $\sigma_i = \beta \cdot IDF(tok_i)$.

Following our notation from earlier, let $PRF_{ColBERT}(\mathcal{I},\theta)$ denote an instantiation of ColBERT-PRF that performs query embedding expansion on the retrieved documents from an earlier stage. Therefore, the full ColBERT-PRF process applied to a document ranking from ColBERT dense retrieval can be written as:

$$\begin{aligned} Ret_{ColBERT-PRF}(\mathcal{I},k') = \ & Ret_{ColBERT}(\mathcal{I},k') \\ & \overset{R}{\gg} PRF_{ColBERT}(\mathcal{I},\theta) \\ & \overset{Q}{\gg} Ret_{ColBERT}(\mathcal{I},k). \end{aligned} \tag{4}$$

However, in this way, $PRF_{ColBERT}(\mathcal{I},\theta)$ returns not a set of expanded query terms with corresponding weights (as returned by $PRF_{RM3}(\mathcal{I}_{ext},\theta)$), but instead a set of expanded query embeddings, with corresponding weights.



(a) Logical flow of ColBERT-PRF for external expansion.

(b) Logical flow of ANCE-PRF for external expansion

Fig. 1. Dense PRF variants for external expansion.

3.4. ANCE dense retrieval

ANCE builds upon a BERT-based dual-encoder (a combined query and document encoder), such that the parameters of the query and the document encoders of ANCE are shared. ANCE encodes a query q using query encoder as ψ_q which can be described as:

$$\psi_q = Encoder_Q([CLS] \oplus q \oplus [SEP]), \tag{5}$$

and encodes a document d into ψ_d using its document encoder:

$$\psi_d = Encoder_D([CLS] \oplus d \oplus [SEP]), \tag{6}$$

where the $[CLS]$ and $[SEP]$ tokens are special tokens in BERT. As a result, ANCE is a single-representation model.

During training, ANCE uses the standard negative log-likelihood loss computed among the above query representation and the relevant as well as the non-relevant document representations. In addition, ANCE employs an asynchronous training strategy where the hard negatives for training are mined using the current latest checkpoint. During retrieval, the semantic similarly, between a query q and document d is measured using the dot product similarity, as

$$s(q,d) = \psi_q \cdot \psi_d. \tag{7}$$

In particular, ANCE uses a FAISS ANN index to record the document embeddings, and identify the nearest neighbouring document embeddings for each query embedding. However, as there is only one embedding for each passage, less aggressive compression can be used compared with ColBERT. We use $Ret_{ANCE}(\mathcal{I}, k)$ to denote the ANCE retrieval.

3.5. ANCE-PRF

As mentioned above, ANCE uses a dual-encoder structure of ANCE that is shared for both queries and documents. Yu, Xiong et al. (2021) proposed ANCE-PRF, which uses a retrained query encoder (where the document encoder is frozen) to leverage the pseudo-relevance feedback documents. Thus, the training process of ANCE-PRF can be summarised as follows: firstly, instead of encoding the query only uses the query text, ANCE-PRF encodes the query uses the query text as well as the feedback document text using:

$$\psi_q^{prf} = Encoder_Q^{prf}([CLS] \oplus q \oplus [SEP] \oplus d_1 \oplus [SEP] \cdots \oplus d_{k'} \oplus [SEP]). \tag{8}$$

Table 2
Summary of the different retrieval and PRF processes used in this work.

Process	Inputs	Outputs
Retrievers ($Q \rightarrow R$)		
BM25	Query text	Retrieved documents
DPH	Query text	Retrieved documents
ColBERT	Query text or Multi-Rep. Query Embs.	Retrieved documents
ANCE	Query text or Single-Rep. Query Emb.	Retrieved documents
PRF ($R \rightarrow Q$)		
RM3	Query text & Retrieved documents text	Query text w/ weights
Bo1	Query text & Retrieved documents text	Query text w/ weights
ColBERT-PRF	Orig. Query Embs. & Retrieved Document Embs.	Multi-Rep. Query Embeddings w/ weights
ANCE-PRF	Query text & Retrieved documents	Single-Rep. Query embedding

Secondly, ANCE-PRF is trained using the same negative likelihood loss as ANCE but with the refined query representation and the relevant as well as the irrelevant document representations. After training, the inference process of ANCE-PRF consists of three stages:

(a) An initial retrieval of ANCE is conducted, matching passages by the similarity of their single embedded representations with the single embedded representation of the query, which returns a list of k' pseudo-relevance feedback passages.

(b) The ANCE-PRF model takes the original query text together with the text of the obtained feedback passages as input, and produces a refined query representation — again, a single embedding.

(c) Finally, another round of ANCE retrieval is performed using the refined query representation.

Following our notation, let $PRF_{ANCE}(\mathcal{I}, \varphi)$ denote an instantiation of ANCE-PRF to refine the query representation based on the retrieved documents from an earlier ANCE stage. Then the full ANCE-PRF retrieval follows the following process:

$$\begin{aligned} Ret_{ANCE-PRF}(\mathcal{I}, k') = \; & Ret_{ANCE}(\mathcal{I}, k') \\ & \overset{R}{\gg} PRF_{ANCE}(\mathcal{I}, \varphi) \\ & \overset{Q}{\gg} Ret_{ANCE}(\mathcal{I}, k). \end{aligned} \tag{9}$$

Thus, $PRF_{ANCE}(\mathcal{I}, \varphi)$ takes the top k' feedback documents retrieved using $Ret_{ANCE}(\mathcal{I}, k')$ together with the input query text to create a refined embedded query representation. Then the refined query embedding is leveraged as new query representation to perform another round of $Ret_{ANCE}(\mathcal{I}, k)$ to return k documents to a user.

3.6. Summary

Table 2 lists the inputs and the corresponding outputs for the retrievers (top half of the table) and the PRF models (bottom half). For instance, among the retrieval models, the inputs to the sparse retrieval models, namely BM25 and DPH, are different from those of the dense retrieval models, ColBERT and ANCE. In particular, the input query text is encoded into multiple query embeddings for ColBERT and into a single query embedding for ANCE. In addition, for the PRF techniques, we see that there are variations in the inputs and outputs are different among the sparse and dense PRF models. Both RM3 and Bo1 take the query text and pseudo-relevance feedback document text as input and produce an expanded query, i.e. words with weights. On the other hand, ColBERT-PRF takes original query embeddings and the embeddings of the feedback document as input and outputs a list of expansion embeddings with weights. In contrast, the ANCE-PRF's inputs are the text of the query and of the feedback documents and its output is a refined single representation query embedding. In this way, it is clear that the PRF and subsequent retrieval stages are tightly-coupled – they cannot be mismatched; however it is possible to change the retrieval stage preceding PRF – either to a different retrieval model, or even to a different index. In the next section, we discuss such formulations of *external expansion*.

4. Our proposed method: Dense external expansion

External expansion (Diaz & Metzler, 2006), also known as collection enrichment (Kwok & Chan, 1998), is a classical PRF-based technique for improving retrieval effectiveness given corpus. In particular, PRF is known to fail if the quality of the feedback documents are poor (Kwok & Chan, 1998; Peng, He et al., 2009; Peng, Macdonald et al., 2009; Xu et al., 2009). To address this, in external expansion, a separate high-quality *external* index is used for an initial retrieval. By obtaining a pseudo-relevant feedback set from the external index, it is assumed to identify additional and higher quality expansion terms than might be found in a pseudo-relevant feedback set obtained on the *target* corpus. This is particularly useful if the top-ranked documents on the target corpus are homogeneous in their choice of vocabulary, or if there are few documents containing the original query terms. By expanding the query using an external corpus, higher diversity of expansion terms can result in more relevant documents being retrieved in the target corpus. In summary, we hypothesise that external expansion can help to address the domain shift when transferring a dense retrieval model towards a target collection in a zero-shot fashion.

Using the notation introduced in Section 3.1, external expansion using a target index $\mathcal{I}_{target}$ and an index of a higher quality corpus, $\mathcal{I}_{ext}$, can be expressed as follows:

$$Ret_{BM25}(\mathcal{I}_{ext}, k') \overset{R}{\gg} PRF_{RM3}(\mathcal{I}_{ext}, \theta) \overset{Q}{\gg} Ret_{BM25}(\mathcal{I}_{target}, k). \tag{10}$$

Based on this, we propose that the dense PRF models, namely ColBERT-PRF and ANCE-PRF, can be instantiated for external expansion, operating entirely in a dense retrieval space:

$$Ret_{Dense}(\mathcal{I}_{ext}, k') \overset{R}{\gg} PRF_{Dense}(\mathcal{I}_{ext}, \theta) \overset{Q}{\gg} Ret_{Dense}(\mathcal{I}_{target}, k), \tag{11}$$

where the subscript $Dense$ can be instantiated as $ColBERT$ or $ANCE$, considering whether the dense retrieval is performed in multiple representation or single representation embedding spaces.

If instantiated as $ColBERT$, a set of expanded query embeddings are obtained from an index of a high quality external corpus, but are then executed on the index of the target corpus, $\mathcal{I}_{target}$. Fig. 1(a) shows the logical flow of query embeddings for ColBERT-PRF in an external expansion setting. In contrast, when instantiating as $ANCE$, a refined query representation is encoded using the high quality feedback documents from the external corpus. The logical flow of ANCE-PRF for external expansion is depicted in Fig. 1(b).

To the best of our knowledge, this is the first application of external expansion in a dense retrieval space. For both ColBERT-PRF and ANCE-PRF models, external expansion is possible as long as the underlying encoders used for both external and target retrieval – and expansion in the case of ANCE-PRF – are unchanged. This ensures that the embedded query representations generated from the external corpus can be used to retrieve documents on the target corpus.

A natural question that arises is as follows: *If PRF_{Dense} already takes contextualised information from the feedback documents into account (thereby addressing word mismatch and polysemy), why is external PRF_{Dense} necessary?* We answer this by arguing that a high quality external corpus, such as Wikipedia, can contain more broader information about the topic, which results in more diverse and valuable expansion embeddings than the local corpus. Moreover, as the expansion embeddings are contextualised, there is less risk of topic drift, as might occur for a classical term-based feedback approach. Similarly, for ANCE-PRF side, the high quality external corpus contains feedback documents may contain a more diverse description of the query topic, thus could further enrich the embedded query representation. On the other hand, using the notation in Eq. (11), it is clear to see that other "hybrid" formulations are possible. Indeed, while the PRF mechanism is tightly coupled (e.g. RM3 generates term-based queries, suitable for the sparse retrieval models such as BM25; ColBERT-PRF generates embedding queries, suitable for ColBERT),[3] the first-stage retrieval can be varied independently. This allows us to vary the choice of first stage ranker — for instance, sparse or multi-representation dense (i.e. ColBERT) vs. single-representation dense (i.e. ANCE).

Indeed, in considering conventional and external PRF configurations in this manner, we are able to determine (a) the impact of dense retrieval on identifying useful feedback documents that may have minimal lexical match with the original query, (b) the value of an external corpora, (c) the benefit of a dense-based PRF technique.

5. Research questions

This paper focuses on improving the performance of zero-shot dense retrieval models. In particular, we investigate the external pseudo-relevance feedback based on dense retrieval feedback information, where the similarity search conducted based on the contextualised information rather than the statistical information used by traditional sparse retrieval models.

Our experiments first study the effectiveness of applying and extracting terms (i.e. sparse feedback) based on external feedback passages that have been identified by a contextualised retrieval model. Thus, we pose our three research questions as follows:

RQ1: What is the benefit of using dense retrieval in obtaining external feedback documents for sparse retrieval?

In addition, we investigate the performance of the zero-shot dense retrieval models using dense external expansion, in both single representation and multiple dense representation based paradigms. Accordingly, we propose the following second research question.

RQ2: What is the benefit of using dense retrieval, in the form of (a) multiple representation (ColBERT) & (b) single representation (ANCE), in obtaining external feedback documents for dense retrieval?

Next, we investigate the impact of the initial stage retrieval for the dense pseudo-relevance feedback models, as follows:

RQ3: Do the dense pseudo-relevance feedback models require dense retrieval to obtain their feedback documents?

Table 3 summarises the configurations of 1st and 2nd stage retrieval paradigms for the baselines and treatments that are tested for each research question.

6. Experimental setup

In the following, we describe the external and target datasets used in this work in Section 6.1. Then the detailed implementation setup and baselines are discussed in Section 6.2 and Section 6.3. Finally, the metrics used in our experiments are detailed in Section 6.4.

6.1. Datasets

Table 4 describes the detailed collection statics for both target collections and external collection used in this work.

[3] Note that the reverse configurations are not tested — for instance, RM3 returns a bag of weighted terms, which would not have sufficient contextual information to be accurately encoded by ColBERT.

Table 3
Summary of the main configurations for each of the research questions.

Research questions	Baselines		Treatments	
	Stage 1 (Target)	Stage 2 (Target)	Stage 1 (External)	Stage 2 (Target)
RQ1	Sparse	Sparse	Dense or Sparse	Sparse
RQ2	Dense	Dense	Dense	Dense
RQ3	Sparse	Dense	Sparse	Dense

Table 4
Statistics of the test collections used.

Corpus	Type	Task	Nbr. Docs.	Nbr. test queries	Avg. query len.	Avg. D/Q
External corpora						
MSMARCO-Passage	External	Passage retrieval	8.8M	–	–	–
Target corpora						
Robust04 title	Target	News retrieval	528k	250	2.76	69.9
Robust04 description	Target	News retrieval	528k	250	15.612	69.9
WT10G title	Target	Web retrieval	1.69M	100	4.23	59.8
WT10G description	Target	Web retrieval	1.69M	100	11.62	59.8
DBPedia	Target	Entity retrieval	4,64M	400	5.39	38.2
NFCorpus	Target	Bio-medical information retrieval	3.6k	50	5.96	38.2
TREC-COVID	Target	Bio-medical information retrieval	171.33k	50	10.60	493.2
Touché-2020	Target	Argument retrieval	382.54k	49	5.39	19.0

Target Collections: In this work, we evaluate using two TREC adhoc test collections, namely the Robust04 and the WT10G. The Robust04 collection uses a corpus of 528k newswire articles from TREC disks 4 & 5; in contrast, WT10G is a collection of 1.69M web documents. For evaluation, we use the title-only and also the (longer) description queries, 250 from Robust04 and 100 from WT10G. TREC provides human labels (relevance assessments) for these queries — on average, there are 69.9 and 59.8 relevant documents for each query of Robust04 and WT10G, respectively.

In addition, we evaluate on four BEIR (Thakur et al., 2021) datasets, namely DBPedia (Hasibi et al., 2017), NFCorpus (Boteva, Gholipour, Sokolov, & Riezler, 2016), TREC-COVID (Voorhees et al., 2021) and Touché-2020 (Bondarenko et al., 2020). As pseudo-relevance feedback techniques are known to be not effective on test collections with few judged documents (Amati, Carpineto, & Romano, 2004), we only evaluate on the BEIR datasets that have a good number of judgements for each query.

External Collection: The external collection used in this work is the MSMARCO passage corpus,[4] which contains approximate 8.8M passages. Indeed, past work (Peng, He et al., 2009; Xu et al., 2009) have successfully used Wikipedia as an external corpus, due to its encyclopedic nature. We use MSMARCO, as it contains high-quality passages about a number of topics, including from Wikipedia (Nguyen et al., 2016).[5]

6.2. Implementation and settings

All the experiments in this work are conducted on the PyTerrier IR experimentation platform (Macdonald & Tonellotto, 2020). We build the ColBERT dense indices following the settings in the original ColBERT (Khattab & Zaharia, 2020) paper, such as padding queries upto a length of 32 tokens, and truncating passages to 180 tokens. We employ the ColBERT model checkpoint and the implementation of the ColBERT-PRF model (Wang et al., 2021) made available by the authors (which was trained on MSMARCO passage training dataset), and follow the ColBERT-PRF default parameter settings reported in Wang et al. (2021), including using 3 feedback passages and 10 expansion embeddings.[6] In addition, we build the ANCE dense indices using the checkpoint released by the authors (Xiong et al., 2021). Similar to ColBERT, ANCE as well as the ANCE-PRF model have also been trained using the MSMARCO passage training dataset. For ANCE-PRF, we experiment with the author (Yu, Xiong et al., 2021) provided model checkpoint trained with 3 feedback passages.[7] Thus, we also use 3 feedback passages for the ANCE-PRF experiments. For all external expansion experiments, we mix the source from external and target corpus. For Robust04 and WT10G, we apply passaging (applying a sliding window of length 150 tokens and stride 75), and documents are ranked by applying max passage. ANCE dense indices are created using model checkpoints released by the original ANCE authors (Xiong et al., 2021), which were also trained on the MSMARCO passage task. For sparse PRF models, we employ both RM3 (Abdul-Jaleel et al., 2004) and Bo1 (Amati & Van Rijsbergen, 2002) techniques and follow the default parameter settings of PyTerrier – i.e. 3 feedback passages and 10 expansion terms. For sparse

[4] https://microsoft.github.io/msmarco/.

[5] Our initial experiments found that performing external expansion using a number of other corpora did not improve retrieval effectiveness on MSMARCO.

[6] https://github.com/terrierteam/pyterrier_colbert.

[7] https://github.com/yuhongqian/ANCE-PRF.

retrieval, we use the stemmed sparse index built using PyTerrier. Our virtual appendix[8] contains the result files of all experiments, and the notebooks needed to reproduce these experiments.

6.3. Baselines

To test the effectiveness of our external pseudo-relevance expansion technique approach, we compare with the following baselines:

- *Sparse Approaches:* We apply sparse retrieval models without PRF, namely the BM25 and DPH weighting models. We also combine these models with sparse pseudo-relevance expansion technique on the target inverted index, i.e. BM25 with RM3 PRF (Abdul-Jaleel et al., 2004) & DPH with Bo1 PRF (Amati & Van Rijsbergen, 2002). Furthermore, we also instantiate these PRF models in an external expansion setting, i.e. $Ret_{BM25}(ext) \overset{R}{\gg} PRF_{RM3}(ext) \overset{Q}{\gg} Ret_{BM25}(target)$ & $Ret_{DPH}(ext) \overset{R}{\gg} PRF_{Bo1}(ext) \overset{Q}{\gg} Ret_{BM25}(target)$.
- *Neural Reranking Approaches:* To aid in our comparisons, we further apply neural rerankers, namely ColBERT and ANCE reranking models upon the sparse retrieval models. For instance, applying a final ColBERT reranker upon BM25 with RM3 query expansion would be denoted as: $Ret_{BM25} \overset{R}{\gg} PRF_{RM3} \overset{Q}{\gg} Ret_{BM25} \overset{R}{\gg} ColBERT$.
- *Dense Approaches:* We also deploy dense retrieval models, with and without a pseudo-relevance feedback mechanism. In particular, for single representation dense retrieval, we deploy ANCE (Xiong et al., 2021), and for multiple representation dense retrieval, we use the ColBERT-E2E (Khattab & Zaharia, 2020) model. For dense retrieval, we apply ColBERT-PRF (Wang et al., 2021) on both normal and external expansion setting.

6.4. Evaluation metrics

We measure effectiveness for the scenarios described in Section 6.1 in terms of Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR), as well as the normalised discounted cumulative gain calculated to rank depth 10 and 20. Moreover, for each model, we compare the Recall calculated to rank depth 1000. When comparing with baseline models, we use the paired t-test ($p < 0.05$) and apply the Holm–Bonferroni multiple testing correction, as per best practices in information retrieval (Sakai, 2021).

7. Results

We address the posed research questions in Sections 7.1–7.3. We summarise the overall findings in Section 7.4, and in Section 7.5 provide an analysis of how ColBERT-PRF performs matching.

7.1. RQ1: Dense external expansion for sparse retrieval

Firstly, we examine the effectiveness of the obtained external feedback for sparse retrieval. In Table 5, for four query sets (Robust04 title-only, Robust04 description-only, WT10G title-only and WT10G description-only), we report results when external retrieval is performed using sparse models (BM25 & DPH), as well using dense retrieval (ANCE & ColBERT). In the table, columns are used to show the processes for each retrieval stage. On each test collection, we firstly report the performance of the four sparse retrieval baselines, including the BM25, BM25 with RM3 query expansion model, DPH, and DPH with Bo1 query expansion model. Then, we measure the retrieval effectiveness of the traditional external expansion models on the sparse retrieval and the external expansion models for sparse retrieval but with feedback documents obtained using dense retrieval.

On analysing Table 5, firstly, we compare the performance of the external expansion models on sparse retrieval using RM3 and Bo1 query expansion models with the sparse retrieval models without any PRF mechanism applied and the sparse PRF models applied only on the target collection on both Robust04 and WT10G test query sets. We notice that both RM3 and Bo1 based external expansion models on sparse retrieval can significantly improve over the models without the external expansion technique applied, which attest the usefulness of the MSMARCO passage as an external collection for both the Robust04 and WT10G corpora. Secondly, we examine the performance of the external expansion with passages produced by dense retrieval. There are four models reported under this external expansion scheme. We observe that the highest values for most of the metrics reported are given by the external expansion models. Similarly, all the external expansion models with dense retrieved passages are significantly improved over all the baselines without external expansion technique applied. When comparing the external expansion models combining dense retrieval with sparse PRF, we find that the former models can significantly improve over the traditional external expansion models using sparse retrieval, which indicates the superiority of the feedback documents produced by dense retrieval viz. the feedback documents produced by the sparse retrieval, e.g. compared to baselines (e) & (f) in Table 5.

Across the query sets, we note that among the external expansion models using dense retrieval, for the title query type, ColBERT is more effective than ANCE, while for the description query type, no obvious pattern emerges among the single and multiple

[8] https://github.com/Xiao0728/DenseExternalExpansion_VirtualAppendix.

Table 5
External expansion for sparse retrieval. The top half table presents the results for Robust04 query sets and the bottom half table presents the results for WT10G query sets. Superscripts a–f denote significant improvements (paired t-test with Holm–Bonferroni correction, $p < 0.05$) over the indicated baseline model(s). The highest value for a queryset is boldfaced.

	1st $\overset{R}{\gg}$	PRF $\overset{Q}{\gg}$	2nd	Robust04 (T)					Robust04 (D)				
				MAP	NDCG@10	NDCG@20	MRR	Recall	MAP	NDCG@10	NDCG@20	MRR	Recall
Target retrieval only													
(a)	BM25	–	–	.241	.432	.406	.654	.687	.245	.437	.411	.680	.690
(b)	BM25$_{target}$	RM3$_{target}$	BM25$_{target}$	.275	.447	.429	.641	.738	.276	.444	.422	.625	.738
(c)	DPH	–	–	.250	.449	.421	.670	.698	.231	.430	.400	.696	.669
(d)	DPH$_{target}$	Bo1$_{target}$	DPH$_{target}$	**.284**	.462	.443	.654	**.752**	.277	.468	.440	.689	.756
External expansion: Sparse external sparse retrieval													
(e)	BM25$_{ext}$	RM3$_{ext}$	BM25$_{target}$	.270^{ac}	.446^{a}	.429^{a}	.632	.731^{ac}	.282^{acd}	.464^{abc}	.436^{abc}	.651^{b}	.748^{ac}
(f)	DPH$_{ext}$	Bo1$_{ext}$	DPH$_{target}$	.278^{ac}	.460^{a}	.438^{ac}	.661^{b}	.740^{ac}	.281^{ac}	.470^{abc}	.439^{abc}	.680^{b}	.750^{ac}
External expansion: Dense external sparse retrieval (Ours)													
	ANCE$_{ext}$	RM3$_{ext}$	BM25$_{target}$	.267^{ab}	.468^{ae}	.440^{ac}	.697^{bef}	.728^{ac}	**.295**^{ac}	. **501**^{abce}	**.470**^{abce}	**.752**^{bef}	**.767**^{ac}
	ANCE$_{ext}$	Bo1$_{ext}$	DPH$_{target}$	.274^{ac}	.467^{a}	.443^{ac}	.715^{abcdef}	.736^{ac}	.280^{ac}	.490^{abcef}	.456^{abcef}	.743^{bef}	.737^{ac}
	ColBERT$_{ext}$	RM3$_{ext}$	BM25$_{target}$	.270^{ac}	.471^{ae}	.441^{a}	.700^{bef}	.729^{ac}	.287^{acd}	.489^{abe}	.458^{abce}	.722^{be}	.759^{ac}
	ColBERT$_{ext}$	Bo1$_{ext}$	DPH$_{target}$	.277^{ac}	**.482**^{acef}	**.455**^{acdef}	**.723**^{abcdef}	.734^{ac}	.280^{acd}	.485^{abce}	.453^{abc}	.738^{abef}	.742^{ac}
				WT10G (T)					WT10G (D)				
Target retrieval only													
(a)	BM25	–	–	.189	.324	.323	.555	.693	.182	.343	.333	.614	.677
(b)	BM25$_{target}$	RM3$_{target}$	BM25$_{target}$	.202	.328	.326	.505	.711	.218	.379	.358	.584	.751
(c)	DPH	–	–	.206	.342	.333	.576	.698	.187	.358	.334	.604	.670
(d)	DPH$_{target}$	Bo1$_{target}$	DPH$_{target}$	.235	.364	.364	.574	.752	.230	.369	.358	.622	.748
External expansion: Sparse external sparse retrieval													
(e)	BM25$_{ext}$	RM3$_{ext}$	BM25$_{target}$	.209^{ac}	.354^{b}	.341	.543^{b}	.729^{abc}	.233^{abc}	.391^{ab}	.383	.620^{abc}	.761^{ac}
(f)	DPH$_{ext}$	Bo1$_{ext}$	DPH$_{target}$	.236^{cbcd}	.383^{abc}	.368^{abc}	.606^{b}	.740^{abc}	.250^{abc}	.405^{abc}	.398^{abcd}	.666^{b}	**.786**^{acd}
External expansion: Dense external sparse retrieval (Ours)													
	ANCE$_{ext}$	RM3$_{ext}$	BM25$_{target}$	.238^{abce}	**.420**^{abcef}	.395^{abce}	.711^{abcdef}	.753^{a}	.251^{abc}	**.534**^{abcde}	.420^{abcde}	.707^{abcde}	.779^{ac}
	ANCE$_{ext}$	Bo1$_{ext}$	DPH$_{target}$	.249^{abce}	.418^{abcef}	.400^{abcdef}	**.721**^{abcdef}	**.782**^{abc}	.252^{abc}	.425^{abcde}	.409^{abcd}	.683^{bce}	.783^{acd}
	ColBERT$_{ext}$	RM3$_{ext}$	BM25$_{target}$	.242^{abce}	.418^{abcef}	.400^{abcef}	.695^{abcdef}	.727^{a}	**.257**^{abce}	.444^{abcde}	**.428**^{abcde}	**.721**^{abcde}	.773^{ac}
	ColBERT$_{ext}$	Bo1$_{ext}$	DPH$_{target}$	**.255**^{abcef}	.417^{abce}	**.402**^{abcdef}	.687^{abcdef}	.768^{abc}	.256^{abc}	.436^{abcde}	.415^{abcde}	.686^{bce}	.776^{acd}

dense retrieval models. This suggests that ColBERT is more suitable for the title (keyword) queries, perhaps due to its token-level embeddings, rather than the single embedding of ANCE.

Overall, in answer to RQ1, we observe that external expansion models supplied with feedback documents obtained from dense retrieval models can bring more benefits for title only queries.

7.2. RQ2: Dense external expansion for dense retrieval

Next, we analyse the effectiveness of external expansion using both ColBERT-PRF and ANCE-PRF, in Sections 7.2.1 and 7.2.2, respectively.

7.2.1. RQ2(a): Dense expansion on multiple representation dense retrieval

We now analyse the performance of the external expansion for dense retrieval models, where the pseudo-relevance feedback information is obtained using dense retrieval models then followed by the ColBERT-PRF contextualised expansion technique. Table 6 reports the results of the external expansion dense retrieval models as well as the sparse query expansion models, the dense retrieval model without any query reformulation techniques applied and the dense retrieval models with ColBERT-PRF applied.

From Table 6, we observe that the dense external expansion models give the highest value for all the metrics on both Robust04 and WT10G title and description query sets. Indeed, ColBERT-PRF improves over ColBERT end-to-end, verifying the results of Wang et al. (2021) on these smaller document corpora. We also find that ColBERT end-to-end dense retrieval model exhibits lower performance than the two sparse query expansion models on both the Robust04 and WT10G query sets – this may indicate underfitting for the title-only (keyword) and description query formulations of Robust04 and WT10G, which differs from the "question-style" of the MSMARCO passage dataset used to train the ColBERT model.

Table 6
External expansion for multiple representation dense retrieval. The top half table presents the results for Robust04 query sets and the bottom half table presents the results for WT10G query sets. Superscripts a–f denote significant improvements (paired t-test with Holm–Bonferroni correction, $p < 0.05$) over the indicated baseline model(s). The highest value for a queryset is boldfaced.

1st $\overset{R}{»}$	PRF $\overset{Q}{»}$	2nd	Robust04 (T)					Robust04 (D)				
			MAP	NDCG@10	NDCG@20	MRR	Recall	MAP	NDCG@10	NDCG@20	MRR	Recall
Baseline runs												
(a) BM25_{target}	RM3_{target}	BM25_{target}	.275	.447	.429	.641	.738	.277	.444	.422	.625	.738
(b) DPH_{target}	Bo1_{target}	DPH_{target}	.284	.462	.443	.654	**.752**	.277	.468	.440	.689	**.756**
(c) ColBERT_{target}	–	–	.237	.447	.421	.701	.608	.220	.435	.401	.685	.605
(d) ColBERT_{target}	$\text{ColBERT-PRF}_{target}$	ColBERT_{target}	.273	.467	.450	.684	.677	.265	.461	.437	.668	.672
(e) BM25_{target}	RM3_{target}	BM25_{target} » ColBERT	.261	.463	.442	.714	.741	.257	.460	.424	.706	.741
(f) DPH_{target}	Bo1_{target}	DPH_{target} » ColBERT	.260	.458	.437	.716	.755	.254	.458	.425	.709	.749
External expansion: Dense external dense retrieval (Ours)												
ColBERT_{ext}	ColBERT-PRF_{ext}	ColBERT_{target}	**.287**^{cd}	**.477**^{ac}	**.467**^{cd}	**.706**^{d}	.714^{cd}	**.281**^{abcd}	**.486**^{c}	**.459**^{c}	**.708**^{a}	.705^{cd}
External expansion: Sparse external dense retrieval (Ours)												
BM25_{ext}	ColBERT-PRF_{ext}	ColBERT_{target}	.241	.429	.411	.634	.669^{c}	.231	.424	.397	.626	.644^{c}
DPH_{ext}	ColBERT-PRF_{ext}	ColBERT_{target}	.243	.431	.414	.677	.674^{c}	.217	.400	.374	.614	.631^{c}
			WT10G (T)					WT10G (D)				
Baseline runs												
(a) BM25_{target}	RM3_{target}	BM25_{target}	.202	.328	.326	.505	.711	.218	.379	.358	.584	**.751**
(b) DPH_{target}	Bo1_{target}	DPH_{target}	**.235**	.364	.364	.574	**.752**	**.230**	.369	.358	.622	.748
(c) ColBERT_{target}	–	–	.160	.356	.337	.614	.510	.162	.360	.339	**.655**	.551
(d) ColBERT_{target}	$\text{ColBERT-PRF}_{target}$	ColBERT_{target}	.183	**.397**	**.372**	.600	.547	.190	.393	.363	.601	.516
(e) BM25_{target}	RM3_{target}	BM25_{target} » ColBERT	.199	.377	.358	.605	.711	.204	.388	.364	.692	.751
(f) DPH_{target}	Bo1_{target}	DPH_{target} » ColBERT	.202	.373	.355	.605	.757	.204	.389	.363	.692	.748
External expansion: Dense external dense retrieval (Ours)												
ColBERT_{ext}	ColBERT-PRF_{ext}	ColBERT_{target}	.216^{cd}	**.397**^{ac}	**.372**^{ac}	**.651**^{a}	.614^{cd}	.226^{cd}	**.408**^{ac}	**.394**^{c}	.651	.644^{cd}
External expansion: Sparse external dense retrieval (Ours)												
BM25_{ext}	ColBERT-PRF_{ext}	ColBERT_{target}	.177	352	.337	.573	.590^{cd}	.199	.366	.360	.603	.645^{cd}
DPH_{ext}	ColBERT-PRF_{ext}	ColBERT_{target}	.178	.360	.348	.568	.594^{cd}	.183	.345	.337	.605	.615^{cd}

Next, we note that external expansion can significantly improve over all the dense baselines across all the metrics and significantly improve over sparse query expansion models, i.e. (a) and (b) baselines in Table 6, in terms of NDCG for title-only queries and MAP for the description queries. One interesting observation is that, although Recall obtained by applying external expansion using ColBERT-PRF outperforms that of both ColBERT end-to-end and ColBERT-PRF performed on the target corpus, it is still lower than the sparse expansion methods. Indeed, this explains the popular practice of applying a more expensive reranker on top of the sparse retrieval models rather than on top of the dense expansion models. Moreover, when we looking back to compare the external expansion dense model with the external expansion sparse model but using dense retrieved passage models in Table 5, we find that some latter model variants exhibit superior performance over the pure dense retrieval based external expansion model. The complementary effect of the contextualised matching models and the statistical information based matching models explains this observation, which is also observed in other recent work (Arabzadeh, Yan, & Clarke, 2021; Gao et al., 2020).

In addition, we further conduct the zero-shot evaluation of the external expansion dense models on four BEIR benchmarks and show the results in Table 7. Firstly, from the top half of Table 7, we find that performing query reformulation using ColBERT-PRF on the target dataset can improve the retrieval effectiveness on all compared datasets except DBPedia. In addition, performing the external dense expansion using ColBERT can further bring significant improvements in terms of NDCG@10 performance on both NFCorpus and Touché-2020 datasets. The ineffectiveness of dense external expansion on the TREC-COVID and NFCorpus datasets are probably due to the external corpora employed (MSMARCO), which contains less biomedical related content. Indeed, TREC-COVID and NFCorpus are biomedical corpora (see Table 4), while MSMARCO is a more general corpus. Moreover, MSMARCO predates the Covid-19 pandemic, and hence is not a good source of external expansion for the TREC-COVID corpus.

In response to RQ2(a), we find that external feedback documents obtained using dense retrieval are beneficial for both external expansion for both sparse & dense feedback and retrieval models. Applying external expansion using a dense retrieval model can significantly improve over the dense and sparse PRF models, i.e. the (a), (b) and (d) baselines in Table 6.

To visualise the impact of the external expansion, Table 8 lists three example queries from the Robust04 title and description query sets. For each query example, we show both the sparse expansion tokens generated by RM3 and the the most likely expansion tokens of the expansion embeddings selected by the ColBERT-PRF model in the target corpora and the external corpora (The FAISS

Table 7
Zero-shot performance in terms of NDCG@10 on BEIR (Thakur et al., 2021). Superscripts a and b denote significant improvements (paired t-test with Holm–Bonferroni correction, $p < 0.05$) over the indicated baseline model(s). The highest NDCG@10 score on a given dataset is boldfaced. W/L denotes the number of queries of our dense external expansion models improved/degraded in terms of the NDCG@10 score of the ColBERT or ANCE model on a given dataset.

Dataset	Normal	Target PRF		External expansion	
	(a) ColBERT_{target}	(b) $\text{ColBERT}_{target} \overset{R}{\gg} \text{ColBERT-PRF}_{target} \overset{Q}{\gg} \text{ColBERT}_{target}$	(W/L)	$\text{ColBERT}_{ext} \overset{R}{\gg} \text{ColBERT-PRF}_{ext} \overset{Q}{\gg} \text{ColBERT}_{target}$ (ours)	(W/L)
DBPedia	**.392**	.387	(121/202)	.353	(154/180)
NFCorpus	.316	.321	(116/87)	**.332**^{ab}	(119/81)
T-COVID	.533	**.548**	(26/18)	.507	(24/23)
Touché-2020	.307	.348	(33/13)	**.353**^{a}	(32/16)
	(a) ANCE_{target}	(b) $\text{ANCE}_{target} \overset{R}{\gg} \text{ANCE-PRF}_{target} \overset{Q}{\gg} \text{ANCE}_{target}$	(W/L)	$\text{ANCE}_{ext} \overset{R}{\gg} \text{ANCE-PRF}_{ext} \overset{Q}{\gg} \text{ANCE}_{target}$ (ours)	(W/L)
DBPedia	.265	.268	(132/137)	**.292**^{ab}	(179/106)
NFCorpus	.236	.239	(86/83)	**.258**^{ab}	(104/63)
T-COVID	.392	**.430**	(28/18)	**.430**	(27/19)
Touché-2020	.291	.292	(27/18)	**.296**	(27/18)

ANN index is used to map an embedding back to the most likely token).[9] The colour of the token indicates the usefulness of the expansion token, with darker indicating higher utility. Usefulness is measured by the difference in Average Precision when that expansion token is removed. From the table, it can be observed that expansion using the external corpus can produce some more useful expansion tokens than the target corpora. For instance, for the query: ‘how are oscar winners selected’, expansion embeddings close to the embedding of ‘glamour’, ‘voting’ and ‘actors’ are selected which can be seen to better identify relevant documents than the target collection (c.f. ‘saturday’, ‘don’). Later, in Section 7.5, we examine the extent that these expansion embeddings match exactly or inexactly with tokens in the documents (i.e. *semantic matches*).

7.2.2. Rq2(b): dense expansion on single representation dense retrieval

We now analyse external dense expansion where the pseudo-relevance feedback documents are produced by the ANCE model, then provided as input for the ANCE-PRF to refine the query representation. Table 9 presents the performance of these configurations, as well as the baselines. Firstly, we see that among the dense retrieval models, performing ANCE-PRF on both Robust04 and WT10G target collections improves over the zero-shot ANCE dense retrieval, i.e baseline ‘d’ outperforms baseline ‘c’ in both collections in (Table 9). We also observe that the external dense retrieval model achieves the highest performance among the three dense retrieval models on all metrics and significantly improve over ANCE performed only on the target collection.

When comparing to the baselines, we notice that the performance of all the zero-shot dense retrieval models on all query sets is lower than sparse expansion models. However, refining the query representation using the pseudo-relevance feedback information of the local collection helps to improve the zero-shot retrieval performance. Performing ANCE-PRF augmentation using the pseudo-relevance feedback documents from a high-quality external corpus results in further improvement. This indicates that the refined query representation from external dense retrieval encapsulates more broad knowledge from the external collection to represent the query. Moreover, compared with the sparse expansion models followed with the ANCE reranker baselines, we notice that applying the ANCE reranker degrades the performance in terms of MAP, NDCG@10 and NDCG@20. This indicates that there is still a large gap in performing the semantic search based on the single representation to the lexical matching.

Moreover, Table 7 presents the zero-shot performance evaluation of dense expansion on single representation dense retrieval model on BEIR benchmarks. From the bottom part of Table 7, we find that dense external expansion using ANCE-PRF exhibits the highest NDCG@10 performance on all the four compared datasets and significantly outperform both the ColBERT and ColBERT-PRF models that are performed entirely on the target datasets. This indicates the usefulness of the external expansion for effective zero-shot evaluation on different benchmarks. This is because ANCE-PRF uses a supervised way of implementing the pseudo-relevance feedback and building upon the large pre-trained BERT model. Thus, the large pre-trained BERT model is capable of generating medical-related knowledge while performing the query refinement even without relevant PRF information as input. While highly effective for ColBERT-PRF, as it performs the PRF technique in an unsupervised way, it might be sensitive to the quality of the external corpus. If there is no relevant context provided by the external corpus, ColBERT-PRF cannot create the relevant expansion embeddings out of thin air.

Overall, in response to RQ2(b), we find that external dense expansion on single representation dense retrieval helps to improve zero-shot dense retrieval performance for both Robust04 and WT10G.

[9] Note that this mapping from embedding to BERT WordPiece token is inexact, hence some apparently meaningless tokens, such as ##up, could actually be a useful expansion *embedding* for retrieval.

Table 8
Qualitative analysis: Examples of the expansion tokens generated by the sparse, namely RM3, and dense PRF, namely ColBERT-PRF, models on the target collection and the external collection for the Robust04 title and description query sets. Expansion tokens (the selected expansion tokens for RM3, or the most likely token for a given expansion embedding for ColBERT-PRF) with higher usefulness are highlighted in darker colour. (The printed version of this article may lack colours - for interpretation of this table, the reader is referred to the web version of this article.)

Robust04 title queries	
Original query terms	308: implant dentistry
Sparse expansion terms (Target)	colleg chiropract implant prosthesi dentistri dental patient dentist devic 1987
Sparse expansion terms (External)	offer dental gener implant teeth tooth dentistri cosmet jaw whiten
Dense expansion tokens (Target)	implant tooth settlement insurance products ##rs life million sales ##1
Dense expansion tokens (External)	jaws dentistry titanium fuse ##′ implant dental tooth ##com replacement
Original query terms	632: southeast asia tin mining
Sparse expansion terms (Target)	burmes burma deleg mine southeast asia myanmar prime command gen
Sparse expansion terms (External)	or mine countri tin southeast china indonesia east asia hemisphe
Dense expansion tokens (Target)	burma cart tin mining vo followed defense where general 000
Dense expansion tokens (External)	bolivia indonesia cass ##′ tin england northern times world also
Original query terms	636: jury duty exemptions
Sparse expansion terms (Target)	serv , man , eslick , exempt , summon , duti , command , juri , murder , soldier .
Sparse expansion terms (External)	2 juri, guidelin , servic , excus , exempt , employe , receiv , duti , court , year .
Dense expansion tokens (Target)	coating , soldier , jury , murder , fees , ##la , regulatory , city , ##2 , we .
Dense xpansion tokens (External)	summons , correspondence , ##ror , jury , ##′ , circuit , duty , bar , five , business .
Robust04 description queries	
Original query terms	633: what is the history of the welsh devolution movement
Sparse expansion terms (Target)	movement vote labour nationalist assembl northern scottish histori elect devolut
Sparse expansion terms (External)	assembl british richard govern wale welsh scottish devolut histori report
Dense expansion tokens (Target)	wales poll ##16 38 won put against cent 000 or
Dense expansion tokens (External)	odds dev literary ##′ scotland ##ol taken community history should
Original query terms	671: find documents that cite the specific benefits the salvation army provides those in need
Sparse expansion terms (Target)	document , find , armi , ford , chariti , benefit , salvat , bush , shop , cite .
Sparse expansion terms (External)	document , includ , contact , armi , salvat , assist , specif , benefit , nearest .
Dense expansion tokens (Target)	northern , valley , se , support , ##ly , many , should , we , who , one .
Dense Expansion tokens (External)	salvation , accepts , rehabilitation , ##′ , documentation , ##ible , army , serving , 24 , health .
Original query terms	685: how are oscar winners selected
Sparse expansion terms (Target)	award box academi pictur select best oscar nomin film winner
Sparse expansion terms (External)	select white best award gown academi actress worn oscar
Dense expansion tokens (Target)	emmy nominations film saturday don ##40 ” ” has as
Dense expansion tokens (External)	glamour winners voting oscar actors ##′ ##up film members actually

7.3. RQ3: Sparse-obtained external feedback for dense retrieval

Besides the dense pseudo-relevance feedback retrieval based on the dense external retrieval as the first stage, we further investigate the performance of the external expansion on the sparse retrieval scenarios. More specifically, we study two external sparse retrieval using BM25 and DPH as the initial stage for both single and multiple representation dense-PRF paradigms.

Table 6 presents the results of the sparse external dense retrieval followed by the ColBERT-PRF query expansion and ColBERT retrieval for both Robust04 and WT10G. Firstly, compared with the baselines runs, we see that using either BM25 or DPH as initial stage retrieval models lead significantly improvement over ColBERT E2E and ColBERT-PRF but underperform the ColBERT-PRF model on the target on other metrics as well as other baseline runs. This demonstrates that sparse external expansion is benefit for retrieving more relevant documents. Secondly, compared with the dense external dense retrieval models, sparse external dense

Table 9
External expansion for single representation dense retrieval. The top half table presents the results for Robust04 query sets and the bottom half table presents the results for WT10G query sets. Superscripts a-f denote significant improvements (paired t-test with Holm–Bonferroni correction, $p < 0.05$) over the indicated baseline model(s). The highest value for a queryset is boldfaced.

	1st $\overset{R}{»}$	PRF $\overset{Q}{»}$	2nd	Robust04 (T)					Robust04 (D)				
				MAP	NDCG@10	NDCG@20	MRR	Recall	MAP	NDCG@10	NDCG@20	MRR	Recall
Baseline runs													
(a)	BM25$_{target}$	RM3$_{target}$	BM25$_{target}$	.275	.447	.429	.641	.738	**.277**	.444	.422	.625	.738
(b)	DPH$_{target}$	Bo1$_{target}$	DPH$_{target}$	**.284**	**.462**	**.443**	**.654**	.752	**.277**	**.468**	**.440**	.689	.756
(c)	ANCE$_{target}$	–	–	.131	.324	.295	.578	.539	.156	.369	.331	.641	.576
(d)	ANCE$_{target}$	ANCE-PRF$_{target}$	ANCE$_{target}$	.155	.345	.312	.586	.541	.165	.381	.343	.649	.563
(e)	BM25$_{target}$	RM3$_{target}$	BM25$_{target}$ » ANCE	.193	.388	.363	.646	.741	.214	.421	.386	.702	.741
(f)	DPH$_{target}$	Bo1$_{target}$	DPH$_{target}$ » ANCE	.193	.384	.359	.643	**.755**	.215	.428	.391	**.712**	**.749**
External expansion: Dense external dense retrieval (Ours)													
	ANCE$_{ext}$	ANCE-PRF$_{ext}$	ANCE$_{target}$	.180^{cd}	.388^{cd}	.354^{cd}	.610^{cd}	.585^{cd}	.183	.403	.368	.665	.585
External expansion: Sparse external dense retrieval (Ours)													
	BM25$_{ext}$	ANCE-PRF$_{ext}$	ANCE$_{target}$	.178^{cd}	.388^{c}	.357^{cd}	.630	.547^{d}	.181	.399	.362	.658	.551
	DPH$_{ext}$	ANCE-PRF$_{ext}$	ANCE$_{target}$	.175^{cd}	.381^{c}	.350^{c}	.616	.550^{d}	.172	.370	.336	.618	.528
				WT10G (T)					WT10G (D)				
Baseline runs													
(a)	BM25$_{target}$	RM3$_{target}$	BM25$_{target}$	.202	.328	.326	.505	.711	.218	**.379**	**.358**	.584	**.751**
(b)	DPH$_{target}$	Bo1$_{target}$	DPH$_{target}$	**.235**	**.364**	**.364**	**.574**	**.752**	**.230**	.369	**.358**	**.622**	.748
(c)	ANCE$_{target}$	–	–	.081	.224	.196	.452	.404	.110	.283	.256	.606	.453
(d)	ANCE$_{target}$	ANCE-PRF$_{target}$	ANCE$_{target}$	.103	.266	.240	.491	.434	.108	.289	.257	.589	.457
(e)	BM25$_{target}$	RM3$_{target}$	BM25$_{target}$ » ANCE	.172	.328	.311	.579	.711	.192	.363	.346	.640	.751
(f)	DPH$_{target}$	Bo1$_{target}$	DPH$_{target}$ » ANCE	.175	.321	.304	.568	.757	.193	.363	.352	.657	.748
External expansion: Dense external dense retrieval (Ours)													
	ANCE$_{ext}$	ANCE-PRF$_{ext}$	ANCE$_{target}$	.117^{cd}	.289^{c}	.259^{c}	.541^{cd}	.452^{c}	.131^{cd}	.314	.291^{cd}	.629	.518^{cd}
External expansion: Sparse external dense retrieval (Ours)													
	BM25$_{ext}$	ANCE-PRF$_{ext}$	ANCE$_{target}$	.123^{c}	.293^{c}	.267^{c}	.580^{c}	.427	.139^{d}	.312	.289	.571	.473
	DPH$_{ext}$	ANCE-PRF$_{ext}$	ANCE$_{target}$	.112^{c}	.276^{c}	.254^{c}	.496	.413	.124	.294	.271	.517	.468

retrieval model exhibit lower performance. These observations are consistent for both Robust04 and WT10G experiments. This indicates that in an end-to-end ColBERT-PRF retrieval paradigm, the dense retrieval is more useful than the sparse retrieval as the first stage to produce high quality pseudo-relevance feedback documents.

Moreover, in an single-representation based ANCE-PRF scenario, Table 9 shows the performance of sparse external dense retrieval models for both Robust04 and WT10G datasets. Firstly, we analyse the top-half table for Robust04. We make the following observations for both the sparse external ANCE-PRF dense retrieval models: (1) they exhibit higher performance than both ANCE (row (e)) and ANCE-PRF (row (f)) performed only on target collection; (2) they show slightly lower performance compared with ANCE reranking models in row (e) and row (f); (3) they show similar performance with the dense external dense retrieval models. On the half-bottom table, we make the following observations for the sparse external dense retrieval models as follows: (1) similar to Robust04, both models outperform the ANCE and ANCE-PRF on WT10G target collection; (2) however, they show a large drop compared with the ANCE reranking models in row (e) and (f); (3) different to the observation for Robust04, sparse external dense retrieval exhibits higher performance than dense external dense retrieval models. Based on this, we find that for the single-representation ANCE-PRF dense retrieval model, sparse external retrieval as the first stage could also produce high quality feedback documents to refine the query representation using ANCE-PRF model.

Thus, in response to RQ3, we find that sparse external retrieval as the initial ranking stage is not sufficient to improve the performance of a multiple representation-based ColBERT-PRF dense retrieval model. However, for the single-representation ANCE-PRF dense retrieval model, sparse external first stage retrieval can improve the retrieval performance over the ANCE baseline, although the ANCE baseline is comparatively weak (emphasising the difficulty of zero-shot single representation dense retrieval).

7.4. Summary of observations

We now report a summary of the observations from the above experiments.

Dense External Expansion, Sparse Retrieval - Section 7.1: We observe that external sparse expansion exhibit a similar performance to the target expansion sparse retrieval models. Moreover, external dense expansion can bring significant improvements over sparse

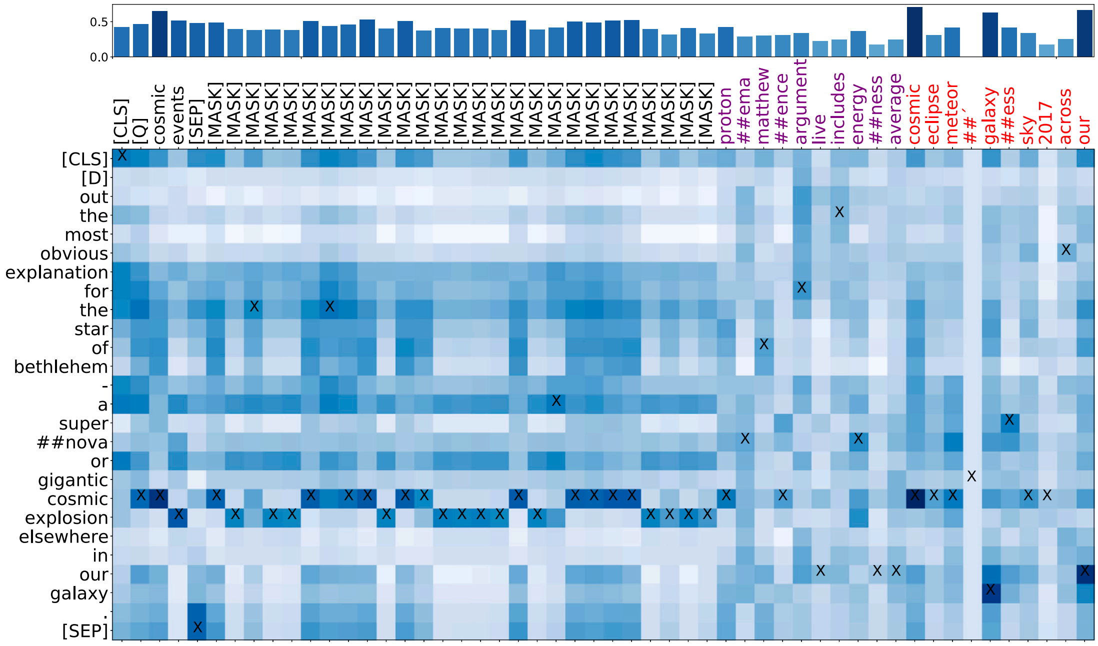


Fig. 2. ColBERT-PRF interaction matrix between Robust04 topic (qid: 405) and document (docid: FT944-864) in an external expansion scenario. The darker shading indicates a higher similarity. The highest similarity among all the documents embeddings for a given query embedding is highlight with a '×' symbol. The top histogram presents the magnitude of contribution for each query embedding to the final score of the document. The expansion tokens generated from the target index are highlight in purple colour while the expansion tokens generated the external index are highlight in red colour. (The printed version of this article may lack colours - for interpretation of this figure, the reader is referred to the web version of this article.)

retrieval models with expansion only performed on the target (12% improvement for Robust04 in NDCG@10: 0.432 → 0.482 and 28% for WT10G: 0.324 → 0.420 in Table 5).

Dense External, Dense Retrieval - Section 7.2: We find that external expansion using ColBERT can significantly improve over the dense retrieval models as well as the dense retrieval with target query expansion (7% improvement for Robust04 in NDCG@10: 0.447 → 0.477 and 21% 0.328 → 0.397 in Table 6); Similarly, external expansion using ANCE can improve the retrieval effectiveness of the dense retrieval (by 20% for Robust04 on NDCG@10: 0.324 → 0.388 and by 29% for WT10G: 0.224 → 0.289 in Table 9). In addition, performing dense external expansion using ColBERT or ANCE can result in further improvements on four BEIR datasets in Table 7.

Sparse External, Dense Retrieval - Section 7.3: We find that sparse external expansion brings limited useful information for ColBERT to improve the followed up dense retrieval effectiveness, and that even applying sparse external retrieval as initial stage can bring useful feedback documents to improve over the dense retrieval on target collection. This emphasises the continuing utility of external expansion in general, even for modern retrieval models.

Overall Performances: The highest NDCG@10 performances observed for Robust04 are 0.482 for title queries and 0.490 for description queries performing external expansion using ColBERT for sparse retrieval (see Table 5). The highest NDCG@10 performances for WT10G are 0.420 for title queries and 0.534 for description queries, obtained by performing external expansion using ANCE for sparse retrieval. Finally, the overall baseline dense retrieval results are not as effective as sparse retrieval in these zero-shot settings, which emphasises the overall difficulty of zero-shot dense retrieval. However, use of dense external expansion can significantly improve effectiveness (e.g. for ColBERT-PRF, NDCG@10 0.447→0.477 in Table 6), it can achieve similar performance to the best sparse retrieval (e.g. 0.482 in Table 5 is not statistically distinguishable from 0.477). This demonstrates the benefit of external expansion for effective zero-shot dense retrieval. In the next section, we analyse to explain the effectiveness of ColBERT-PRF.

7.5. ColBERT-PRF semantic matching analysis

We now analyse the extent that ColBERT-PRF prefers exact matches versus inexact (semantic) matching. Indeed, as the ColBERT model's matching behaviour is performed using contextualised BERT embeddings for each token, polysemous words (which have the same surface form, but different meanings) will have distinct embeddings, while synonymous words will have similar embeddings. Hence it is possible to see the extent that synonymous words, or more generally, semantic matches – where the query token is not matched with an embedding representing the same word in the document – are occurring during retrieval. Indeed, we see this as a particular advantage for the ColBERT multiple representation model, which is not possible for models such as ANCE where the whole query and the whole document are each represented in a single representation embedding. In short, we further investigate

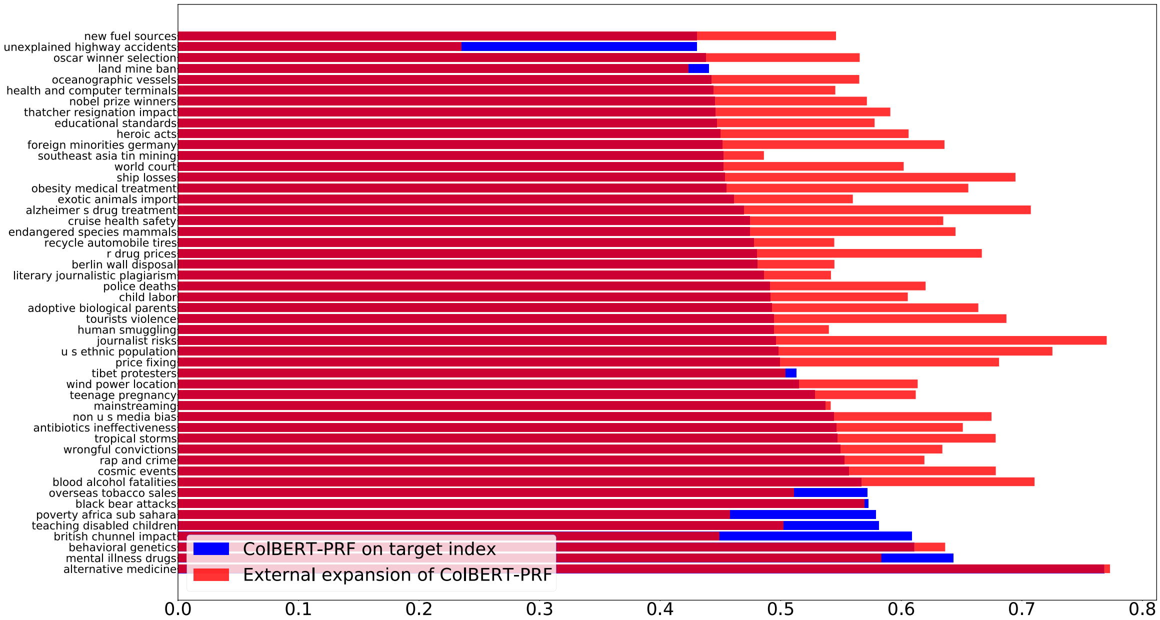
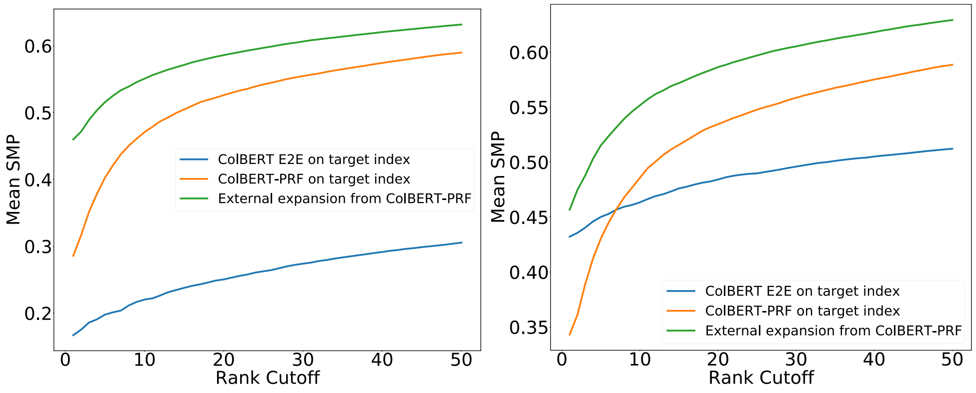


Fig. 3. Per-query semantic matching proportion for 50 of the Robust04 title topics. (The printed version of this article may lack colours - for interpretation of this figure, the reader is referred to the web version of this article.)



(a) Mean SMP on Robust04 title queries

(b) Mean SMP on Robust04 description queries

Fig. 4. Mean semantic matching proportion (Mean SMP) as rank varies. (The printed version of this article may lack colours - for interpretation of this figure, the reader is referred to the web version of this article.)

the extent that semantic matching occurs when performing query expansion using PRF documents generated from an high quality external collection compared to only in the target collection.

Fig. 2 depicts the interaction matrix of ColBERT-PRF model in an external expansion scenario between the Robust04 title topic: “cosmic events” and its top ranked document. Across the top, the original query tokens, along with the ‘[MASK]’ query embeddings added for “query augmentation” (see Section 3.2) are in black; the (most likely) tokens identified by ColBERT-PRF from the target corpus are shown in purple, and those identified by ColBERT-PRF from the external corpus are in red. On analysis of the figure, we observe that for the original query token ‘cosmic’ experiences exact matching as the token is in the same form with its corresponding returned highest Max-Sim scored document token. In contrast, the original query token ‘events’ experiences semantic matching, as its corresponding document token with the highest Max-Sim score is ‘explosion’. Indeed, this match is semantically contextualised in nature, in that it is unlikely that ‘events’ would be somehow matched ‘explosion’ except in the context of ‘cosmic’. Its also possible to see the ‘[MASK]’ query embeddings added by ColBERT are mostly similar in nature to the original query terms ‘cosmic’ and ‘events’, by virtue of the fact their Max-Sim matches are with the same document tokens. When looking at the tokens for the expansion embeddings, we see that the target index generates expansion embeddings seem unrelated to the query (e.g. ‘matthew’). In contrast, the tokens for the externally-sources expansion embeddings are more related in nature to the query (‘eclipse’, ‘meteor’, ‘galaxy’). Of these, ‘galaxy’ experiences an exact match, while other expanded embeddings experience semantic matching (e.g. ‘eclipse’ matches with ‘cosmic’.

To quantify the extent that semantic matching takes place, we define a new measure that inspects the Max-Sim, and determines whether each query embedding is matched with the same token (exact match) vs. an inexact (semantic) match with a different token. Formally, let t_i and t_j respectively denote the token id of the ith query embedding and jth document embedding, respectively. Given a query q and the top ranked k documents, R_k, we define the *Semantic Match Proportion* (SMP) at rank cutoff k w.r.t. q and R_k as:

$$SMP(q, R_k) = \sum_{d \in R_k} \frac{\sum_{i \in \mathsf{toks}(q)} \mathbb{1}[t_i \neq t_j] \cdot \max_{j=1,\ldots,|d|} \phi_{q_i}^T \phi_{d_j}}{\sum_{i \in \mathsf{toks}(q)} \max_{j=1,\ldots,|d|} \phi_{q_i}^T \phi_{d_j}}, \tag{12}$$

where $\mathsf{toks}(q)$ returns the indices of the query embeddings that correspond to BERT tokens, i.e., not [CLS], [Q], or [MASK] tokens,[10] R_k is the top ranked k documents, and $\mathbb{1}[]$ is the indicator function.

Now, we measure the difference of the semantic matching proportion performances with and without applying the external expansion from the multiple representation dense retrieval following Eq. (12). Fig. 3 depicts the per-query semantic matching proportion for the first 50 Robust04 title queries against the top 1 document for the external expansion of ColBERT-PRF and ColBERT-PRF on the target index. We observe that among the sampled queries, 41/50 queries’ semantic matching are increased when performing ColBERT-PRF using feedback documents from the external collection rather than the target. This indicates that the pseudo-relevance feedback documents generated from the external collection contain broader and more useful information than the small target collection to refine the original query representation to get closer to the relevant document representations in the semantic matching space.

Next, we investigate the Mean SMP of different approaches, namely the ColBERT E2E on the target index, the ColBERT-PRF on the target index and the external expansion from ColBERT-PRF, at different rank cutoffs, k. In particular, Fig. 4 depicts the observed Mean SMP on both the Robust04 title-only and description-only query sets. On observing Fig. 4(a), we find that both of the ColBERT-PRF based approaches exhibit higher mean semantic matching proportion than the ColBERT E2E approach performed on different rank cutoffs. In addition, external expansion from ColBERT-PRF shows higher Mean SMP than ColBERT-PRF applied using only the target index. Indeed, as most of the Robust04 title-only queries are short queries, their information needs cannot be sufficiently described only using the original query representation. This observation further verifies the findings from Fig. 3. Expanding the ColBERT query representations with expansion embeddings results in a better query representation, while expanding from a higher quality external corpus will further improve the representation. Next, from Fig. 4(b), we observe that for the description queries, at the initial ranks, ColBERT E2E exhibits higher semantic matching than the ColBERT-PRF on the target index. Put another way, ColBERT-PRF tend to experience higher exact matching on the very high ranks than ColBERT E2E. However, across the range of rank cutoff values, external expansion from ColBERT-PRF show the highest semantic matching compared to both ColBERT-PRF and ColBERT E2E on the target collection.

Overall, we find that the external expansion for multiple representation dense retrieval result in higher semantic matching proportion than expansion only performed in the target index. This demonstrates further the value of performing pseudo-relevance feedback using ColBERT-PRF.

8. Conclusions

This work has revisited the pseudo-relevance feedback, in the form of external expansion, as applied to improve the zero-shot retrieval. In particular, our experiments employed popular dense retrieval models from both the single representation and multiple representation families to extract useful feedback documents from high-quality external corpus (MSMARCO). More specifically, we investigate different frameworks performing external expansion for zero-shot retrieval and conduct extensive experiments on two TREC test collections (Robust04 and WT10G) and four BEIR datasets (DBPedia, NFCorpus, TREC-COVID and Touché-2020), namely (a) *dense external expansion for sparse retrieval*, (b) *dense external expansion for dense retrieval* and (c) *sparse-obtained external feedback for dense retrieval*. Overall, we find that high quality feedback documents obtained from both multiple representation dense retrieval and single representation dense retrieval can significantly improve sparse retrieval on both test collections (by 12% and 28% for Robust04 and WT10G, respectively). Moreover, we observed that performing external dense expansion can significantly outperform the zero-shot dense retrieval models on target collection. In addition, we find that pseudo-relevance feedback documents produced by sparse retrieval model are benefit to augment query representation. Finally, we thoroughly investigate the semantic matching analysis for ColBERT-PRF and observe that performing external expansion using multiple representation dense retrieval results in higher semantic matching proportion than performing on the target. In future work, we plan to investigate the selective application of a dense PRF mechanism on external and target corpora, to address those queries that benefit most from external expansion.

CRediT authorship contribution statement

Xiao Wang: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualisation, Writing – original draft. **Craig Macdonald:** Writing – review & editing, Project administration, Investigation, Validation, Supervision, Methodology, Software, Resources, Conceptualisation. **Iadh Ounis:** Writing – review & editing, Project administration, Investigation, Supervision, Methodology, Conceptualisation.

[10] Indeed, [CLS], [Q], and [MASK] do not correspond to actual WordPiece tokens originating from the user’s query and hence can never have exact matches, so we exclude them from this calculation.

Acknowledgements

Xiao Wang acknowledges support by the China Scholarship Council (CSC) from the Ministry of Education of P.R. China. Craig Macdonald and Iadh Ounis acknowledge EPSRC, United Kingdom grant EP/R018634/1: Closed-Loop Data Science for Complex, Computationally- & Data-Intensive Analytics.

References

Abdul-Jaleel, Nasreen, Allan, James, Croft, W Bruce, Diaz, Fernando, Larkey, Leah, Li, Xiaoyan, et al. (2004). UMass at TREC 2004: Novelty and HARD. In *Proceedings of TREC*.

Amati, Gianni, Carpineto, Claudio, & Romano, Giovanni (2004). Query difficulty, robustness, and selective application of query expansion. In *Proceedings of ECIR* (pp. 127–137).

Amati, Gianni, & Van Rijsbergen, Cornelis Joost (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 357–389.

Arabzadeh, Negar, Yan, Xinyi, & Clarke, Charles L. A. (2021). Predicting efficiency/effectiveness trade-offs for dense vs. Sparse retrieval strategy selection. In *Proceedings of CIKM* (pp. 2862–2866).

Azad, Hiteshwar Kumar, & Deepak, Akshay (2019). Query expansion techniques for information retrieval: a survey. *Information Processing & Management*, *56*(5), 1698–1735.

Bondarenko, Alexander, Fröbe, Maik, Beloucif, Meriem, Gienapp, Lukas, Ajjour, Yamen, Panchenko, Alexander, et al. (2020). Overview of touché 2020: argument retrieval. In *Proceddings of CLEF* (pp. 384–395).

Boteva, Vera, Gholipour, Demian, Sokolov, Artem, & Riezler, Stefan (2016). A full-text learning to rank dataset for medical information retrieval. In *Proceddings of ECIR* (pp. 716–722).

Chen, Xiaoyang, Hui, Kai, He, Ben, Han, Xianpei, Sun, Le, & Ye, Zheng (2022). Incorporating ranking context for end-to-end BERT Re-ranking. In *Proceedings of ECIR* (pp. 111–127). Springer.

Chen, Tao, Zhang, Mingyang, Lu, Jing, Bendersky, Michael, & Najork, Marc (2022). Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. In *Proceedings of ECIR*.

Croft, W. Bruce, Metzler, Donald, & Strohman, Trevor (2010). *Search engines: Information retrieval in practice, Vol. 520*. Addison-Wesley Reading.

Dai, Zhuyun, & Callan, Jamie (2020). Context-aware document term weighting for ad-hoc search. In *Proceedings of WWW* (pp. 1897–1907).

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of ACL* (pp. 4171–4186).

Diaz, Fernando, & Metzler, Donald (2006). Improving the estimation of relevance models using large external corpora. In *Proceedings of SIGIR* (pp. 154–161).

Formal, Thibault, Piwowarski, Benjamin, & Clinchant, Stéphane (2021). SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of SIGIR* (pp. 2288–2292).

Gao, Luyu, Dai, Zhuyun, Chen, Tongfei, Fan, Zhen, Van Durme, Benjamin, & Callan, Jamie (2020). Complementing lexical retrieval with semantic residual embedding. In *Proceedings of ECIR* (pp. 146–160).

Hasibi, Faegheh, Nikolaev, Fedor, Xiong, Chenyan, Balog, Krisztian, Bratsberg, Svein Erik, Kotov, Alexander, et al. (2017). DBpedia-entity v2: a test collection for entity search. In *Proceedings of SIGIR* (pp. 1265–1268).

He, Ben, & Ounis, Iadh (2007). Combining fields for query expansion and adaptive query expansion. *Information Processing & Management*, *43*(5), 1294–1307.

Johnson, Jeff, Douze, Matthijs, & Jégou, Hervé (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547.

Karpukhin, Vladimir, Oguz, Barlas, Min, Sewon, Lewis, Patrick, Wu, Ledell, Edunov, Sergey, et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP* (pp. 6769–6781).

Khattab, Omar, & Zaharia, Matei (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of SIGIR* (pp. 39–48).

Kwok, Kui Lam, & Chan, Margaret (1998). Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of SIGIR* (pp. 250–256).

Lavrenko, Victor, & Croft, W. Bruce (2001). Relevance based language models. In *Proceedings of SIGIR* (pp. 120–127).

Li, Hang, Mourad, Ahmed, Zhuang, Shengyao, Koopman, Bevan, & Zuccon, Guido (2021). Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. *ACM Transactions on Information Systems (TOIS)*.

Li, Canjia, Sun, Yingfei, He, Ben, Wang, Le, Hui, Kai, Yates, Andrew, et al. (2018). NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval. In *Proceedings of EMNLP* (pp. 4482–4491).

Li, Hang, Zhuang, Shengyao, Mourad, Ahmed, Ma, Xueguang, Lin, Jimmy, & Zuccon, Guido (2021). Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study. In *Proceedings of ECIR* (pp. 599–612).

Lin, Jimmy, Nogueira, Rodrigo, & Yates, Andrew (2021). Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies*, *14*(4), 1–325.

Lioma, Christina, & Ounis, Iadh (2008). A syntactically-based query reformulation technique for information retrieval. *Information Processing & Management*, *44*(1), 143–162.

MacAvaney, Sean, Cohan, Arman, & Goharian, Nazli (2020). SLEDGE-Z: A zero-shot baseline for COVID-19 literature search. In *Proceedings of EMNLP* (pp. 4171–4179).

MacAvaney, Sean, Nardini, Franco Maria, Perego, Raffaele, Tonellotto, Nicola, Goharian, Nazli, & Frieder, Ophir (2020). Expansion via prediction of importance with contextualization. In *Proceedings of SIGIR* (pp. 1573–1576).

MacAvaney, Sean, Yates, Andrew, Cohan, Arman, & Goharian, Nazli (2019). CEDR: Contextualized embeddings for document ranking. In *Proceedings of SIGIR* (pp. 1101–1104).

Macdonald, Craig, & Tonellotto, Nicola (2020). Declarative experimentation in information retrieval using PyTerrier. In *Proceedings of ICTIR* (pp. 161–168).

Macdonald, Craig, Tonellotto, Nicola, & Ounis, Iadh (2021). On single and multiple representations in dense passage retrieval. In *IIR 2021 workshop*.

Mallia, Antonio, Khattab, Omar, Suel, Torsten, & Tonellotto, Nicola (2021). Learning passage impacts for inverted indexes. In *Proceedings of SIGIR* (pp. 1723–1727).

Naseri, Shahrzad, Dalton, Jeffrey, Yates, Andrew, & Allan, James (2021). CEQE: Contextualized embeddings for query expansion. In *Proceedings of ECIR*.

Nguyen, Tri, Rosenberg, Mir, Song, Xia, Gao, Jianfeng, Tiwary, Saurabh, Majumder, Rangan, et al. (2016). MS MARCO: A Human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Nogueira, Rodrigo, Lin, Jimmy, & Epistemic, A. I. (2019). From doc2query to docTTTTTquery. *Online Preprint*.

Nogueira, Rodrigo, Yang, Wei, Lin, Jimmy, & Cho, Kyunghyun (2019). Document expansion by query prediction. arXiv preprint arXiv:1904.08375.

Pan, Min, Wang, Junmei, Huang, Jimmy X, Huang, Angela J, Chen, Qi, & Chen, Jinguang (2022). A probabilistic framework for integrating sentence-level semantics via BERT into pseudo-relevance feedback. *Information Processing & Management*, *59*(1), Article 102734.

Peng, Jie, He, Ben, & Ounis, Iadh (2009). Predicting the usefulness of collection enrichment for enterprise search. In *Proceedings of CIKM* (pp. 366–370).

Peng, Jie, Macdonald, Craig, He, Ben, & Ounis, Iadh (2009). A study of selective collection enrichment for enterprise search. In *Proceedings of CIKM* (pp. 1999–2002).

Rocchio, Joseph (1971). Relevance feedback in information retrieval. *The Smart Retrieval System-Experiments in Automatic Document Processing*, 313–323.

Sakai, Tetsuya (2021). On Fuhr's guideline for IR evaluation. *SIGIR Forum*, *54*(1).

Thakur, Nandan, Reimers, Nils, Rücklé, Andreas, Srivastava, Abhishek, & Gurevych, Iryna (2021). BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of NeurIPS*.

Voorhees, Ellen (2005). The TREC robust retrieval track. In *ACM SIGIR forum, Vol. 39* (pp. 11–20).

Voorhees, Ellen (2006). The TREC 2005 robust track. In *ACM SIGIR forum, Vol. 40* (pp. 41–48).

Voorhees, Ellen, Alam, Tasmeer, Bedrick, Steven, Demner-Fushman, Dina, Hersh, William R, Lo, Kyle, et al. (2021). TREC-COVID: Constructing a pandemic information retrieval test collection. In *ACM SIGIR forum* (pp. 1–12).

Wang, Le, Luo, Ze, Li, Canjia, He, Ben, Sun, Le, Yu, Hao, et al. (2020). An end-to-end pseudo relevance feedback framework for neural document retrieval. *Information Processing & Management*, *57*(2), Article 102182.

Wang, Xiao, Macdonald, Craig, & Tonellotto, Nicola (2021). Pseudo-relevance feedback for multiple representation dense retrieval. In *Proceedings of ICTIR* (pp. 297–306).

Wang, Junmei, Pan, Min, He, Tingting, Huang, Xiang, Wang, Xueyan, & Tu, Xinhui (2020). A pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval. *Information Processing & Management*, *57*(6), Article 102342.

Wong, WS, Luk, Robert Wing Pong, Leong, Hong Va, Ho, KS, & Lee, Dik Lun (2008). Re-examining the effects of adding relevance information in a relevance feedback environment. *Information Processing & Management*, *44*(3), 1086–1116.

Xiong, Lee, Xiong, Chenyan, Li, Ye, Tang, Kwok-Fung, Liu, Jialin, Bennett, Paul, et al. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of ICLR*.

Xu, Yang, Jones, Gareth J. F., & Wang, Bin (2009). Query dependent pseudo-relevance feedback based on Wikipedia. In *Proceedings of SIGIR* (pp. 59–66).

Yu, HongChien, Dai, Zhuyun, & Callan, Jamie (2021). PGT: Pseudo relevance feedback using a graph-based transformer. In *Proceedings of ECIR* (pp. 440–447). Springer.

Yu, HongChien, Xiong, Chenyan, & Callan, Jamie (2021). Improving query representations for dense retrieval with pseudo relevance feedback. In *Proceedings of CIKM* (pp. 3592–3596).

Zheng, Zhi, Hui, Kai, He, Ben, Han, Xianpei, Sun, Le, & Yates, Andrew (2020). BERT-QE: Contextualized query expansion for document re-ranking. In *Proceedings of EMNLP: findings* (pp. 4718–4728).