Do the Findings of Document and Passage Retrieval Generalize to the Retrieval of Responses for Dialogues?

Penha, Gustavo; Hauff, Claudia

DOI
10.1007/978-3-031-28241-6_9

Publication date
2023

Document Version
Final published version

Published in
Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Proceedings

Citation (APA)
Penha, G., & Hauff, C. (2023). Do the Findings of Document and Passage Retrieval Generalize to the Retrieval of Responses for Dialogues? In J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, A. Caputo, & U. Kruschwitz (Eds.), *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Proceedings* (pp. 132-147). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 13982). Springer. https://doi.org/10.1007/978-3-031-28241-6_9

Do the Findings of Document and Passage Retrieval Generalize to the Retrieval of Responses for Dialogues?

Gustavo Penha(✉) and Claudia Hauff

TU Delft, Delft, The Netherlands
{g.penha-1,c.hauff}@tudelft.nl

Abstract. A number of learned sparse and dense retrieval approaches have recently been proposed and proven effective in tasks such as passage retrieval and document retrieval. In this paper we analyze with a replicability study if the lessons learned generalize to the retrieval of responses for dialogues, an important task for the increasingly popular field of conversational search. Unlike passage and document retrieval where documents are usually longer than queries, in response ranking for dialogues the queries (dialogue contexts) are often longer than the documents (responses). Additionally, dialogues have a particular structure, i.e. multiple utterances by different users. With these differences in mind, we here evaluate how generalizable the following major findings from previous works are: **(F1)** query expansion outperforms a no-expansion baseline; **(F2)** document expansion outperforms a no-expansion baseline; **(F3)** zero-shot dense retrieval underperforms sparse baselines; **(F4)** dense retrieval outperforms sparse baselines; **(F5)** hard negative sampling is better than random sampling for training dense models. Our experiments (https://github.com/Guzpenha/transformer_rankers/tree/full_rank_retrieval_dialogues.)—based on three different information-seeking dialogue datasets—reveal that four out of five findings (**F2**–**F5**) generalize to our domain.

1 Introduction

Conversational search is concerned with creating agents that satisfy an information need by means of a *mixed-initiative* conversation through natural language interaction, rather than through the traditional search engine results page. A popular approach to conversational search is retrieval-based [3]: given an ongoing conversation and a large corpus of historic conversations, retrieve the response that is best suited from the corpus [11,28,45,47,48]. Due to the effectiveness of heavily pre-trained transformer-based language models such as BERT [4], they have become the predominant approach for conversation response re-ranking [8,28,42,43,53].

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13982, pp. 132–147, 2023.
https://doi.org/10.1007/978-3-031-28241-6_9

The most common evaluation procedure for conversation response re-ranking consists of re-ranking a limited set of n candidate responses (including the ground-truth response(s)), followed by measuring the number of relevant responses found in the first K positions—$Recall_n@K$ [52]. Since the entire collection of available responses is typically way bigger[1] than such a set of candidates, this setup is in fact a selection problem, where we have to choose the correct response out of a few options. This evaluation overlooks the first-stage retrieval step, which retrieves a set of n responses to be re-ranked. If the first-stage model, e.g. BM25, fails to retrieve relevant responses, the entire pipeline fails.

Motivated by a lack of research on the first-stage retrieval step, we are interested in answering in our replicability study whether the considerable knowledge obtained on document and passage retrieval tasks generalizes to the dialogue domain. Unlike document and passage retrieval where the documents are generally longer than the queries, in response retrieval for dialogues the queries (dialogue contexts) tend to be longer than the documents (responses). A second important difference is the structure induced by the dialogue as seen in Table 1.

Table 1. Comparison between passage retrieval and response retrieval for dialogues. In Sect. 3 we define the task of *First-stage Retrieval for Dialogues*. p^+/r^+ are the relevant passage/response.

	Passage retrieval	First-stage retrieval for dialogues
Input	Query q	Dialogue context $\mathcal{U} = \{u^1, u^2, ..., u^\tau\}$
Example	q: *what is theraderm used for*	u^1: *I was in the mood to play Chrono Trigger again [...] Is there a performant SNES emulator that has that feature?* u^2: {url} *allows you to map joypad buttons to keyboard keys and [...]* u^3: *Do the diagonals for the analog stick work correctly for you? [...]*
Output	Ranked list of passages	Ranked list of responses
Example	p^+: *Thera-Derm Lotion is used as a moisturizer to treat [...]*	r^+: *In the"Others" tab, try [...]*

Given the differences between the domains, we verify empirically across three information-seeking datasets and 1.7M queries, the generalizability of five findings (**F1** to **F5**) from the passage and document retrieval literature related to state-of-the-art sparse and dense retrieval models. We are motivated in our selection of these five findings by their impact in prior works (cf. Sect. 2). Our results show that four out of five previous findings do indeed generalize to our domain:

[1] While for most benchmarks [52] we have only 10–100 candidates, a working system with the Reddit data from PolyAI https://github.com/PolyAI-LDN/conversational-datasets would need to retrieve from 3.7 billion responses.

F1 ✗[2] Dialogue context (i.e. query) expansion outperforms a no-expansion baseline [1,18,21,49].

F2 ✓ Response (i.e. document) expansion outperforms a no-expansion baseline [19,21,25] *if the expansion model is trained to generate the most recent context (last utterance*[3] *of the dialogue) instead of older context (all utterances).*

F3 ✓ Dense retrieval in the zero-shot[4] setting underperforms sparse baselines [34,41] *except when it goes through intermediate training on large amounts of out-of-domain data.*

F4 ✓ Dense retrieval with access to target data[5] outperforms sparse baselines [7,15,34] *if an intermediate training step on out-of-domain data is performed before the fine-tuning on target data.*

F5 ✓ Harder negative sampling techniques lead to effectiveness gains [46,51] *if a denoising technique is used to reduce the number of false negative samples.*

Our results indicate that most findings translate to the domain of retrieval of responses for dialogues. A promising future direction is thus to start with successful models from other domains—for which there are more datasets and previous research—and study how to adapt and improve them for retrieval-based conversational search.

2 Related Work

In this section we first discuss current research in retrieval-based systems for conversational search, followed by reviewing the major findings of (un)supervised sparse and dense retrieval in the domains of passage and document retrieval.

2.1 Ranking and Retrieval of Responses for Dialogues

Early neural models for response re-ranking were based on matching the representations of the concatenated dialogue context and the representation of a response in a single-turn manner with architectures such as CNN and LSTM [14,23]. More complex neural architectures matching each utterance with the response were also explored [9,22,54]. Heavily pre-trained language models such as BERT were first shown to be effective by Nogueira and Cho [24] for re-ranking. Such models quickly became a predominant approach for re-ranking in IR [21] and were later shown to be effective for re-ranking responses in conversations [28,42].

In contrast, the first-stage retrieval of responses for a dialogue received relatively little attention [29]. Lan et al. [17] and Tao et al. [38] showed that BERT-based dense retrieval models outperform BM25 for first-stage retrieval of

[2] ✗ indicates that the finding does not hold in our domain whereas ✓ indicates that it holds in our domain followed by the *necessary condition or exception.*

[3] For example in Table 1 the last utterance is u^3.

[4] A zero-shot is a model that does not have access to target data, cf. Table 2.

[5] Target data is data from the same distribution, i.e. dataset, of the evaluation dataset.

responses for dialogues. A limitation of their work is that strong sparse retrieval baselines that have shown to be effective in other retrieval tasks, e.g. BM25 with *dialogue context expansion* [25] or BM25 with *response expansion* [49], were not employed for dense retrieval. We do such comparisons here and test a total of five major findings that have been not been evaluated before by previous literature on the first-stage retrieval of responses for dialogues.

2.2 Dense and Sparse Models for Passage and Document Retrieval

Context for F1. Retrieval models can be categorized into two dimensions: supervised vs. unsupervised and dense vs. sparse representations [19]. An *unsupervised* sparse representation model such as BM25 [35] represents each document and query with a sparse vector with the dimension of the collection's vocabulary, having many zero weights due to non-occurring terms. Since the weights of each term are entirely based on term statistics they are considered unsupervised methods. Such approaches are prone to the vocabulary mismatch problem [6], as semantic matches are not considered. A way to address such a problem is by using query expansion methods. RM3 [1] is a competitive [49] query expansion technique that uses pseudo-relevance feedback to add new terms to the queries followed by another final retrieval step using the modified query.

Context for F2. A *supervised* sparse retrieval model can take advantage of the effectiveness of transformer-based language models by changing the terms' weights from collection statistics to something that is learned. Document expansion with a learned model can be considered a learned sparse retrieval approach [19]. The core idea is to create pseudo documents that have expanded terms and use them instead when doing retrieval. Doc2query [25] is a strong supervised sparse retrieval baseline that uses a language model to predict queries that might be issued to find a document. The predictions of this model are used to create the augmented pseudo documents.

Context for F3 and F4. Supervised dense retrieval models[6], such as ANCE [46] and coCodenser [7], represent query and documents in a small fixed-length space, for example of 768 dimensions. Dense retrieval models without access to target data for training—known as the *zero-shot scenario*—have underperformed sparse methods (**F3**). For example, the `BEIR` benchmark [41] showed that BM25 was superior to dense retrieval from 9–18 (depending on the model) out of the 18 datasets in the zero-shot scenario. In contrast, when having access to enough supervision from target data, dense retrieval models have shown to consistently outperform strong sparse baselines [7, 15, 34] (**F4**).

[6] A distinction can also be made of cross-encoders and bi-encoders, where the first encode the query and document jointly as opposed to separately [40]. Cross-encoders are applied in a re-ranking step due to their inefficiency and thus are not our focus.

Context for F5. In order to train neural ranking models, a small set of negative (i.e. non-relevant) candidates are necessary as it is prohibitively expensive to use every other document in the collection as negative sample for a query. A limitation of randomly selecting negative samples is that they might be too easy for the ranking model to discriminate from relevant ones, while for negative documents that are harder the model might still struggle. For this reason hard negative sampling has been shown to perform better than random sampling for passage and document retrieval [36,46,51].

3 First-Stage Retrieval for Dialogues

In this section we first describe the problem of first-stage retrieval of responses, followed by the findings we want to replicate from sparse and dense approaches.

Problem Definition. The task of first-stage retrieval of responses for dialogues, concerns retrieving the best response out of the entire collection given the dialogue context. Formally, let $\mathcal{D} = \{(\mathcal{U}_i, \mathcal{R}_i, \mathcal{Y}_i)\}_{i=1}^{M}$ be a data set consisting of M triplets: dialogue context, response candidates and response relevance labels. The dialogue context $\mathcal{U}_i$ is composed of the previous utterances $\{u^1, u^2, ..., u^\tau\}$ at the turn τ of the dialogue. The candidate responses $\mathcal{R}_i = \{r^1, r^2, ..., r^n\}$ are either ground-truth responses r^+ or negative sampled candidates r^-, indicated by the relevance labels $\mathcal{Y}_i = \{y^1, y^2, ..., y^n\}$. In previous work, the number of candidates is limited, typically $n = 10$ [29]. The findings we replicate here come from passage and document retrieval tasks where there is no limit to the number of documents or passages that have to be retrieved. Thus, in all of our first-stage retrieval task experiments n is set to the size of the entire collection of responses in the corpus. The number of ground-truth responses is one, the observed response in the conversational data. The task is then to learn a ranking function $f(.)$ that is able to generate a ranked list from the entire corpus of responses $\mathcal{R}_i$ based on their predicted relevance scores $f(\mathcal{U}, r)$.

F1: Unsupervised Sparse Retrieval. We rely on classic retrieval methods, for which the most commonly used baseline is BM25. One of the limitations of sparse retrieval is the vocabulary mismatch problem. Expansion techniques are able to overcome this problem by appending new words to the dialogue contexts and responses. For this reason, we here translate a query expansion technique to the dialogue domain and perform *dialogue context expansion* with RM3 [1], a competitive unsupervised method that assumes that the top-ranked responses by the sparse retrieval model are relevant. From these pseudo-relevant responses, words are selected and an expanded dialogue context is created and subsequently employed by the sparse retrieval method to rank the final list of responses. **The effectiveness of RM3 in the domain of dialogues is the first finding that we validate**.

F2: Learned Sparse Retrieval. Alternatively, we can expand the responses in the collection with a learned method. To do so we "translate" doc2query [25] into our domain, yielding resp2ctxt. Formally, we fine-tune a generative transformer model G for the task of generating the dialogue context $\mathcal{U}_i$ from the ground-truth response r_i^+. This model is then used to generate expansions for all responses in the collection, $r^i = concat(r^i, G(r^i))$. These expansions are appended to the responses and the collection is indexed again—the sparse retrieval method itself is not modified, i.e. we continue using BM25. This approach (which we coin resp2ctxt) leads to two improvements: term re-weighting (adding terms that already exist in the document) and dealing with the vocabulary mismatch problem (adding new terms). **The effectiveness of doc2query in the domain of dialogues is the second finding that we validate**.

Unlike passage and document retrieval where the queries are smaller than the documents, for the retrieval of responses for dialogues the queries are longer than the documents[7]. This is a challenge for the generative model, since generating larger pieces of text is a more difficult problem than smaller ones as there is more room for errors. Motivated by this, we also explored a modified version of resp2ctxt that aims to generate only the last utterance of the dialogue context: resp2ctxt$_{lu}$. This model is trained to generate u^τ from r_i^+, instead of trying to generate the whole utterance $\mathcal{U}_i = \{u^1, u^2, ..., u^\tau\}$. The underlying premise is that the most important utterance from the dialogue is the last one, and if it is correctly generated by resp2ctxt$_{lu}$, the sparse retrieval method will be able to find the correct response from the collection.

F3: Zero-Shot Dense Retrieval. We rely on methods that learn to represent the dialogue context and the responses separately in a dense embedding space. Responses are then ranked by their similarity to the dialogue context. We rely here on pre-trained language transformer models, such as BERT [4] and MPNet [37], to obtain such representations of the dialogue context and response. This approach is generally referred to as a *bi-encoder* model [21] and is an effective family of models[8]. A zero-shot model is one that is not trained on the target data. Target data is data from the same distribution, i.e. dataset, of the evaluation dataset.

One way of improving the representations of a heavily pre-trained language model for the zero-shot setting is to fine-tune it with intermediate data [33]. Such intermediate data contains triplets of query, relevant document, and negative document and can include multiple datasets. The advantage of adding this step before employing the representations of the language model is to reduce the

[7] For example, while the `TREC-DL-2020` passage and document retrieval tasks the queries have between 5–6 terms on average and the passages and documents have over 50 and 1000 terms respectively, for the information-seeking dialogue datasets used here the dialogue contexts (queries) have between 70 and 474 terms on average depending on the dataset while the responses (documents) have between 11 and 71.

[8] See for example the top models in terms of effectiveness from the MSMarco benchmark leaderboards https://microsoft.github.io/msmarco/.

gap between the pre-training and the downstream task at hand [26,30,31]. In Table 2 we clarify the relationship between pre-training, intermediate training and fine-tuning.

Table 2. The different training stages and data, their purposes, examples of datasets, and the type of dense model obtained after each stage.

	Pre-training data	Intermediate data	Target data
Purpose	Learn general representations	Learn sentence representations for ranking	Learn representations for target distribution
Model is	Zero-shot	Zero-shot	Fine-tuned
Example	`Wikipedia`	`MSMarco`	`MANtIS`

The intermediate training step learns to represent pieces of text (query and documents) by applying a mean pooling function over the transformer's final layer, which is then used to calculate the dot-product similarity. The loss function employs multiple negative texts from the same batch to learn the representations in a constrastive manner, also known as in-batch negative sampling. Such a procedure learns better text representations than a naive approach that uses the $[CLS]$ token representation of BERT [2,33].

The function $f(\mathcal{U}, r)$ is then $dot(\eta(concat(\mathcal{U})), \eta(r))$, where η is the representation obtained by applying the mean pooling function over the last layer of the transformer model, and $concat(\mathcal{U}) = u^1 \mid [U] \mid u^2 \mid [T] \mid ... \mid u^{\tau}$, where $\mid$ indicates the concatenation operation. The utterances from the context $\mathcal{U}$ are concatenated with special separator tokens $[U]$ and $[T]$ indicating end of utterances and turns[9]. **The effectiveness of a zero-shot bi-encoder model in the domain of dialogues is the third finding we validate.**

F4: Fine-Tuned Dense Retrieval. The standard procedure is to fine-tune dense models with target data that comes from the same dataset that the model will be evaluated. Since we do not have labeled negative responses, all the remaining responses in the dataset can be thought of as non-relevant to the dialogue context. Computing the probability of the correct response over all other responses in the dataset would give us $P(r \mid \mathcal{U}) = \frac{P(\mathcal{U},r)}{\sum_k P(\mathcal{U},r_k)}$. This computation is prohibitively expensive, and the standard procedure is to approximate it using a few negative samples. The *negative sampling* task is then as follows: given the dialogue context $\mathcal{U}$ find challenging responses r^- that are non-relevant for $\mathcal{U}$. Negative sampling can be seen as a retrieval task, where one can use a model to retrieve negatives by applying a retrieval function to the collection of responses using $\mathcal{U}$ as the query.

[9] The special tokens $[U]$ and $[T]$ will not have any meaningful representation in the zero-shot setting, but they can be learned on the fine-tuning step.

With such a dataset at hand, we continue the training—after the intermediate step—in the same manner as done by the intermediate training step, with the following cross-entropy loss function[10] for a batch with size B:

$$\mathcal{J}(\mathcal{U}, \mathbf{r}, \theta) = -\frac{1}{B} \sum_{i=1}^{B} \left[f\left(\mathcal{U}_i, r_i\right) - \log \sum_{j=1, j!=i}^{B} e^{f\left(\mathcal{U}_i, r_j\right)} \right],$$

where $f(\mathcal{U}, r)$ is the dot-product of the mean pooling of the last layer of the transformer model. **The effectiveness of a fine-tuned bi-encoder model in the domain of dialogues is the fourth finding we validate here.**

F5: Hard Negative Sampling. A limitation of random samples is that they might be too easy for the ranking model to discriminate from relevant ones, while for negative documents that are hard the model might still struggle. For this reason, another popular approach is to use a ranking model to retrieve negative documents using the given query with a classic retrieval technique such as BM25. This leads to finding negative documents that are closer to the query in the sparse representation space, and thus they are *harder negatives*. Since dense retrieval models have been outperforming sparse retrieval in a number of cases with available training data, more complex negative sampling techniques making use of dense retrieval have also been proposed [12,46]. **The effectiveness of hard negative sampling for a bi-encoder model in the domain of dialogues is the fifth finding we validate here.**

4 Experimental Setup

In order to compare the different sparse and dense approaches we consider three large-scale information-seeking conversation datasets[11]: `MSDialog` [32] contains 246K context-response pairs, built from 35.5K information seeking conversations from the Microsoft Answer community, a QA forum for several Microsoft products; `MANtIS` [27] contains 1.3 million context-response pairs built from conversations of 14 Stack Exchange sites, such as *askubuntu* and *travel*; $\texttt{UDC}_{\texttt{DSTC8}}$ [16] contains 184k context-response pairs of disentangled Ubuntu IRC dialogues.

Implementation Details. For BM25 and BM25+RM3[12] we rely on `pyserini` implementations [20]. In order to train resp2ctxt expansion methods we rely on the Huggingface `transformers` library [44], using the *t5-base* model. We fine-tune the T5 model for 2 epochs, with a learning rate of 2e−5, weight decay of

[10] We refer to this loss as MultipleNegativesRankingLoss.

[11] `MSDialog` is available at https://ciir.cs.umass.edu/downloads/msdialog/; `MANtIS` is available at https://guzpenha.github.io/MANtIS/; $\texttt{UDC}_{\texttt{DSTC8}}$ is available at https://github.com/dstc8-track2/NOESIS-II.

[12] We perform hyperparameter tuning using grid search on the number of expansion terms, number of expansion documents, and weight.

0.01, and batch size of 5. When augmenting the responses with resp2ctxt we follow docT5query [25] and append three different context predictions, using sampling and keeping the top-10 highest probability vocabulary tokens.

For the zero-shot dense models, we rely on the `SentenceTransformers` [33] model releases. The library uses Hugginface's `transformers` for the pre-trained models such as BERT [4] and MPNet [37]. For the bi-encoder models, we use the pre-trained *all-mpnet-base-v2* weights which were the most effective in our initial experiments, compared with other pre-trained models[13]. When fine-tuning the dense retrieval models, we rely on the *MultipleNegativesRankingLoss*, which accepts a number of hard negatives, and also uses the remaining in-batch random negatives to train the model. We use a total of 10 negative samples for dialogue context. We fine-tune the dense models for a total of 10k steps, and every 100 steps we evaluate the models on a re-ranking task that selects the relevant response out of 10 responses. We use the re-ranking validation MAP to select the best model from the whole training to use in evaluation. We use a batch size of 5, with 10% of the training steps as warmup steps. The learning rate is 2e−5 and the weight decay is 0.01. We use `FAISS` [13] to perform the similarity search.

Evaluation. To evaluate the effectiveness of the retrieval systems we use $R@K$. We thus evaluate the models' capacity of finding the correct response out of the whole possible set of responses[14]. We perform Students t-tests at the 0.95 confidence level with Bonferroni correction to compare statistical significance of methods. Comparisons are performed across the results for each dialogue context.

5 Results

In this section, we discuss our empirical results along with the five major findings from previous work (Sect. 1) in turn. Table 3 contains the main results regarding **F1** to **F4**. Table 5 contains the results for **F5**.

F1 ✗ Query expansion via RM3 leads to improvements over not using query expansion [1,18,21,49]. BM25+RM3 (row 1b) does not improve over BM25 (1a) on any of the three conversational datasets analyzed. We performed thorough hyperparameter fine-tuning and no combination of the RM3 hyperparameters outperformed BM25. **This indicates that F1 does not hold for the task of response retrieval for dialogues.**

A manual analysis of the new terms appended to a sample of 60 dialogue contexts by one of the paper's authors revealed that only 18% of them have at least one relevant term added based on our best judgment. Unlike web search

[13] The alternative models we considered are those listed in the model overview section at https://www.sbert.net/docs/pretrained_models.html.

[14] The standard evaluation metric in conversation response ranking [8,39,50] is recall at position K with n candidates $R_n@K$. Since we are focused on the first-stage retrieval we set n to be the entire collection of answers.

Table 3. Results for the generalizability of F1–F4. Bold values indicate the highest recall for each type of approach. Superscripts indicate statistically significant improvements using Students t-test with Bonferroni correction. † = *other methods from the same group* 1 = *best from unsupervised sparse retrieval;* 2 = *best from supervised sparse retrieval;* 3 = *best from zero-shot dense retrieval.* For example, in F3 † indicates that row (3d) improves over rows (3a–c), ^{1} indicates that it improves over row (1a) and ^{2} indicates it improves over row (2b).

		MANtIS		MSDialog		UDC$_{\text{DSTC8}}$	
		R@1	**R@10**	**R@1**	**R@10**	**R@1**	**R@10**
(0)	Random	0.000	0.000	0.000	0.001	0.000	0.001
Unsupervised sparse					**F1**		
(1a)	BM25	**0.133**†	**0.299**†	**0.064**†	**0.177**†	**0.027**†	**0.070**†
(1b)	BM25 + RM3	0.073	0.206	0.035	0.127	0.011	0.049
Supervised sparse					**F2**		
(2a)	BM25 + resp2ctxt	0.135	0.309	0.074	**0.208**	0.028	0.067
(2b)	BM25 + resp2ctxt$_{lu}$	**0.147**$^{\dagger 1}$	**0.325**$^{\dagger 1}$	**0.075**^{1}	0.202^{1}	**0.029**	**0.076**
Zero-shot dense (Model$_{IntermediateData}$)					**F3**		
(3a)	ANCE$_{600K-MSMarco}$	0.048	0.111	0.050	0.124	0.010	0.028
(3b)	TAS-B$_{400K-MSMarco}$	0.062	0.143	0.060	0.157	0.019	0.050
(3c)	Bi-encoder$_{215M-mul}$	0.138	0.297	0.108	0.277	0.023	0.076
(3d)	Bi-encoder$_{1.17B-mul}$	**0.155**$^{\dagger 1}$	**0.341**$^{\dagger 12}$	**0.147**$^{\dagger 12}$	**0.339**$^{\dagger 12}$	**0.041**†	**0.097**$^{\dagger 12}$
Fine-tuned dense (Model$_{NegativeSampler}$)					**F4**		
(4a)	Bi-encoder$_{Random(0)}$	**0.130**	**0.307**	**0.168**^{123}	**0.387**^{123}	**0.050**^{12}	**0.128**^{123}

where the query is often incomplete, under-specified, and ambiguous, in the information-seeking datasets employed here the dialogue context (query) is quite detailed and has more terms than the responses (documents). We hypothesize that because the dialogue contexts are already quite descriptive, the task of expansion is trickier in this domain and thus we observe many dialogues for which the added terms are noisy.

F2 ✓ Document expansion via resp2ctxt leads to improvements over no expansion [19,21,25]. We find that a naive approach to response expansion improves marginally in two of the three datasets with BM25+resp2ctxt (2a) outperforming BM25 (1a). However, the proposed modification of predicting only the last utterance of the dialogue (resp2ctxt$_{lu}$) performs better than predicting the whole utterance, as shown by BM25+resp2ctxt$_{lu}$'s (2b) higher recall values. In the MANtIS dataset the R@10 goes from 0.309 when using the model trained to predict the dialogue context to 0.325 when using the one trained to predict only the last utterance of the dialogue context. **We thus find that F2 generalizes to response retrieval for dialogues, especially when predicting only the last utterance of the context**[15]**.**

[15] As future work, more sophisticated techniques can be used to determine which parts of the dialogue context should be predicted.

Table 4. Statistics of the augmentations for resp2ctxt and resp2ctxt$_{lu}$. New words are the ones that did not exist in the document before.

	MANtIS	MSDialog	UDC$_{\text{DSTC8}}$
Context avg length	474.12	426.08	76.95
Response avg length	42.58	71.38	11.06
Aug. avg length - resp2ctxt	494.23	596.99	202.3
Aug. avg length - resp2ctxt$_{lu}$	138.5	135.29	72.57
% new words - resp2ctxt	71%	69%	71%
% new words - resp2ctxt$_{lu}$	59%	37%	63%

In order to understand what the response expansion methods are doing most—term re-weighting or adding novel terms—we present the percentage of novel terms added by both methods in Table 4. The table shows that resp2ctxt$_{lu}$ does more term re-weighting than adding new words when compared to resp2ctxt (53% and 70% on average are new words respectively and thus 47% vs 30% are changing the weights by adding existing words), generating overall smaller augmentations (115.45 vs 431.17 on average respectively).

F3 ✓ Sparse retrieval outperforms zero-shot dense retrieval [34,41]**.** Sparse retrieval models are more effective than the majority of zero-shot dense models, as shown by the comparison of rows (1a–b), and (2a–b) with rows (3a–c). However, a dense retrieval model that has gone through intermediate training on large and diverse datasets including dialogues is more effective than a strong sparse retrieval model, as we see by comparing row (3d) with row (2b) in Table 3.

For example, while the zero-shot dense retrieval models based only on the MSMarco dataset (3a–b) perform on average 35% worse than the strong sparse baseline (2b) in terms of R@10 for the MSDialog dataset, the zero-shot model trained with 1.17B instances on diverse data (3d) is 68% better than the sparse baseline (2b). When using a bigger amount of intermediate training data[16], we see that the zero-shot dense retrieval model (3d) is able to outperform the sparse retrieval baseline by margins of 33% of R@10 on average across datasets.

We thus show that F3 only generalizes to response retrieval for dialogues if we do not employ a large set of diverse intermediate data. As expected, the closer the intermediate training data distribution is to the evaluation data, the better the dense retrieval model performs. The results indicate that a good zero-shot retrieval model needs to go through intermediate training on a large set of training data coming from multiple datasets to generalize well to different domains and outperform strong sparse retrieval baselines.

F4 ✓ Dense models with access to target training data outperform sparse models [7,15,34]**.** First, we see that fine-tuning the dense retrieval

[16] For the full description of the intermediate data see https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

model, which has gone through intermediate training already, with random sampling—row (4a) in Table 3—achieves the best overall effectiveness in two of the three datasets. **This result shows that F4 generalizes to the task of response retrieval for dialogues when employing intermediate training**[17]. Having access to the target data as opposed to only the intermediate training data means that the representations learned by the model are closer to the true distribution of the data.

We hypothesize that fine-tuning the bi-encoder for MANtIS (4a) is harmful because the intermediate data contains Stack Exchange responses. In this way, the set of dialogues of Stack Exchange that MANtIS encompasses might be serving only to overfit the intermediate representations. As evidence for this hypothesis, we found that (I) the learning curves flatten quickly (as opposed to other datasets) and (II) fine-tuning another language model that does not have Stack Exchange data (MSMarco) in their fine-tuning, bi-encoder$_{bert-base}$ (3c), improves the effectiveness with statistical significance from 0.092 R@10 to 0.205 R@10.

F5 ✓ Hard negative sampling is better than random sampling for training dense retrieval models [46, 51]**.** Surprisingly we find that naively using more effective models to select negative candidates is detrimental to the effectiveness of the dense retrieval model (see Hard negative sampling in Table 5). We observe this phenomenon when using different language models, when switching intermediate training on or off for all datasets, and when using an alternative contrastive loss [10] that does not employ in-batch negative sampling[18].

After testing for a number of hypotheses that might explain why harder negatives do not improve the effectiveness of the dense retrieval model, we found that false negative samples increase significantly when using better negative sampling methods. False negatives are responses that are potentially valid for the context. Such relevant responses lead to unlearning relevant matches between context and responses as they receive negative labels. See below an example of a false negative sample retrieved by the bi-encoder model (row 3d of Table 3):

Dialogue context ($\mathcal{U}$): hey... how long until dapper comes out? [U] 14 days [...] [U] i thought it was coming out tonight
Correct response (r^{+}): just kidding couple hours
False negative sample (r^{-}): there is a possibility dapper will be delayed [...] meanwhile, dapper discussions should occur in ubuntu+1

Denoising techniques try to solve this problem by reducing the number of false negatives. We employ a simple approach that instead of using the top-ranked responses as negative responses, we use the bottom responses of the top-ranked responses as negatives[19]. This decreases the chances of obtaining false positives and if $k << |\mathcal{D}|$ we will not obtain random samples. Our experiments in Table 5

[17] Our experiments show that when we do not employ the intermediate training step the fine-tuned dense model does not generalize well, with row (3d) performance dropping to 0.172, 0.308 and 0.063 R@10 for MANtIS, MSDialog and UDC$_{DSTC8}$ respectively.

[18] The results are not shown here due to space limitations.

[19] For example, if we retrieve $k = 100$ responses, instead of using responses from top positions 1–10, we use responses 91–100 from the bottom of the list.

reveal that this denoising technique, row (3b), increases the effectiveness for harder negative samples, beating all models from Table 3 for two of the three datasets. **The results indicate that F5 generalizes to the task of response retrieval for dialogues only when employing a denoising technique.**

Table 5. Results for the generalizability of F5—with and without a denoising strategy for hard negative sampling. Superscripts indicate statistically significant improvements using Students t-test with Bonferroni correction . †*=significance against the random sampling baseline,* ‡*=significance against hard negative sampling without denoising.*

	`MANtIS`	`MSDialog`	`UDC`$_{\texttt{DSTC8}}$
	R@10	R@10	R@10
Baseline			
(1) Bi-encoder$_{Random}$	0.307	0.387	**0.128**
Hard negative sampling			
(2a) Bi-encoder$_{BM25}$	0.271	0.316	0.087
(2b) Bi-encoder$_{Bi-encoder}$	0.146	0.306	0.051
Denoised hard negative sampling			
(3a) Bi-encoder$_{BM25}$	0.257	0.358‡	0.121‡
(3b) Bi-encoder$_{Bi-encoder}$	**0.316**†‡	**0.397**†‡	0.107‡

6 Conclusion

In this work, we tested if the knowledge obtained in dense and sparse retrieval from experiments on the tasks of passage and document retrieval generalizes to the first-stage retrieval of responses for dialogues. Our replicability study reveals that while most findings do generalize to our domain, a simple translation of the models is not always successful. A careful analysis of the domain in question might reveal better ways to adapt techniques.

As future work, we believe an important direction is to evaluate learned sparse methods that do weighting and expansion for both the queries and documents [5]—while resp2ctxt is able to both change the weights of the terms in the response (by repeating existing terms) and expand terms (by adding novel terms), it is not able to do weighting and expansion for the dialogue contexts.

Acknowledgements. This research has been supported by NWO projects SearchX (639.022.722) and NWO Aspasia (015.013.027).

References

1. Abdul-Jaleel, N., et al.: Umass at trec 2004: Novelty and hard. Computer Science Department Faculty Publication Series, p. 189 (2004)
2. Aghajanyan, A., Gupta, A., Shrivastava, A., Chen, X., Zettlemoyer, L., Gupta, S.: Muppet: massive multi-task representations with pre-finetuning. arXiv preprint arXiv:2101.11038 (2021)

3. Anand, A., Cavedon, L., Joho, H., Sanderson, M., Stein, B.: Conversational search (dagstuhl seminar 19461). In: Dagstuhl Reports. vol. 9. Schloss Dagstuhl-Leibniz-Zentrum für Informatik (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Formal, T., Lassance, C., Piwowarski, B., Clinchant, S.: Splade v2: Sparse lexical and expansion model for information retrieval. arXiv preprint arXiv:2109.10086 (2021)
6. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. Commun. ACM **30**(11), 964–971 (1987)
7. Gao, L., Callan, J.: Unsupervised corpus aware language model pre-training for dense passage retrieval. arXiv preprint arXiv:2108.05540 (2021)
8. Gu, J.C., Li, T., Liu, Q., Ling, Z.H., Su, Z., Wei, S., Zhu, X.: Speaker-aware Bert for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 2041–2044 (2020)
9. Gu, J.C., Ling, Z.H., Liu, Q.: Interactive matching network for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2321–2324 (2019)
10. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 2, pp. 1735–1742. IEEE (2006)
11. Han, J., Hong, T., Kim, B., Ko, Y., Seo, J.: Fine-grained post-training for improving retrieval-based dialogue systems. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1549–1558. Association for Computational Linguistics, Online, June 2021. https://doi.org/10.18653/v1/2021.naacl-main.122, https://aclanthology.org/2021.naacl-main.122
12. Hofstätter, S., Lin, S.C., Yang, J.H., Lin, J., Hanbury, A.: Efficiently teaching an effective dense retriever with balanced topic aware sampling. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 113–122 (2021)
13. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2019)
14. Kadlec, R., Schmid, M., Kleindienst, J.: Improved deep learning baselines for ubuntu corpus dialogs. arXiv preprint arXiv:1510.03753 (2015)
15. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)
16. Kummerfeld, J.K., et al.: A large-scale corpus for conversation disentanglement. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/p19-1374, http://dx.doi.org/10.18653/v1/P19-1374
17. Lan, T., Cai, D., Wang, Y., Su, Y., Mao, X.L., Huang, H.: Exploring dense retrieval for dialogue response selection. arXiv preprint arXiv:2110.06612 (2021)
18. Lin, J.: The simplest thing that can possibly work: pseudo-relevance feedback using text classification. arXiv preprint arXiv:1904.08861 (2019)
19. Lin, J.: A proposed conceptual framework for a representational approach to information retrieval. arXiv preprint arXiv:2110.01529 (2021)

20. Lin, J., Ma, X., Lin, S.C., Yang, J.H., Pradeep, R., Nogueira, R.: Pyserini: a Python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021), pp. 2356–2362 (2021)
21. Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: Bert and beyond. In: Synthesis Lectures on Human Language Technologies, vol. 14(4), 1–325 (2021)
22. Lin, Z., Cai, D., Wang, Y., Liu, X., Zheng, H.T., Shi, S.: The world is not binary: Learning to rank with grayscale data for dialogue response selection. arXiv preprint arXiv:2004.02421 (2020)
23. Lowe, R., Pow, N., Serban, I., Pineau, J.: The ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909 (2015)
24. Nogueira, R., Cho, K.: Passage re-ranking with Bert. arXiv preprint arXiv:1901.04085 (2019)
25. Nogueira, R., Lin, J., Epistemic, A.: From doc2query to docttttquery. Online preprint 6 (2019)
26. Peeters, R., Bizer, C., Glavaš, G.: Intermediate training of Bert for product matching. Small **745**(722), 2–112 (2020)
27. Penha, G., Balan, A., Hauff, C.: Introducing mantis: a novel multi-domain information seeking dialogues dataset. arXiv preprint arXiv:1912.04639 (2019)
28. Penha, G., Hauff, C.: Curriculum learning strategies for IR: an empirical study on conversation response ranking. arXiv preprint arXiv:1912.08555 (2019)
29. Penha, G., Hauff, C.: Challenges in the evaluation of conversational search systems. In: Converse@ KDD (2020)
30. Poth, C., Pfeiffer, J., Rücklé, A., Gurevych, I.: What to pre-train on? efficient intermediate task selection. arXiv preprint arXiv:2104.08247 (2021)
31. Pruksachatkun, Y., et al.: Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? arXiv preprint arXiv:2005.00628 (2020)
32. Qu, C., Yang, L., Croft, W.B., Trippas, J.R., Zhang, Y., Qiu, M.: Analyzing and characterizing user intent in information-seeking conversations. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 989–992 (2018)
33. Reimers, N., Gurevych, I.: Sentence-Bert: Sentence embeddings using SIAMESE Bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, November 2019. https://arxiv.org/abs/1908.10084
34. Ren, R., et al.: A thorough examination on zero-shot dense retrieval. arXiv preprint arXiv:2204.12755 (2022)
35. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR 1994, pp. 232–241. Springer, London (1994). doi:https://doi.org/10.1007/978-1-4471-2099-5_24
36. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 (2020)
37. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: MpNet: masked and permuted pretraining for language understanding. Adv. Neural. Inf. Process. Syst. **33**, 16857–16867 (2020)

38. Tao, C., Feng, J., Liu, C., Li, J., Geng, X., Jiang, D.: Building an efficient and effective retrieval-based dialogue system via mutual learning. arXiv preprint arXiv:2110.00159 (2021)
39. Tao, C., Wu, W., Xu, C., Hu, W., Zhao, D., Yan, R.: Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In: WSDM, pp. 267–275 (2019)
40. Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I.: Augmented sbert: data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. arXiv preprint arXiv:2010.08240 (2020)
41. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: Beir: a heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint arXiv:2104.08663 (2021)
42. Whang, T., Lee, D., Lee, C., Yang, K., Oh, D., Lim, H.: An effective domain adaptive post-training method for Bert in response selection. arXiv preprint arXiv:1908.04812 (2019)
43. Whang, T., Lee, D., Oh, D., Lee, C., Han, K., Lee, D.H., Lee, S.: Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14041–14049 (2021)
44. Wolf, T., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
45. Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z.: Sequential matching network: a new architecture for multi-turn response selection in retrieval-based chatbots. In: ACL, pp. 496–505 (2017)
46. Xiong, L., et al.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. arXiv preprint arXiv:2007.00808 (2020)
47. Yang, L., et al.: IART: intent-aware response ranking with transformers in information-seeking conversation systems. arXiv preprint arXiv:2002.00571 (2020)
48. Yang, L., et al.: Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In: SIGIR pp. 245–254 (2018)
49. Yang, W., Lu, K., Yang, P., Lin, J.: Critically examining the "neural hype" weak baselines and the additivity of effectiveness gains from neural ranking models. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1129–1132 (2019)
50. Yuan, C., et al.: Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In: EMNLP, pp. 111–120 (2019)
51. Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., Ma, S.: Optimizing dense retrieval model training with hard negatives. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1503–1512 (2021)
52. Zhang, Z., Zhao, H.: Advances in multi-turn dialogue comprehension: a survey. arXiv preprint arXiv:2110.04984 (2021)
53. Zhang, Z., Zhao, H.: Structural pre-training for dialogue comprehension. arXiv preprint arXiv:2105.10956 (2021)
54. Zhou, X., et al.: Multi-turn response selection for chatbots with deep attention matching network. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1118–1127 (2018)