BERT-based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval

Shuai Wang
The University of Queensland
Brisbane, Australia
shuai.wang5@uqconnect.edu.au

Shengyao Zhuang
The University of Queensland
Brisbane, Australia
shengyao.zhuang@uq.edu.au

Guido Zuccon
The University of Queensland
Brisbane, Australia
g.zuccon@uq.edu.au

ABSTRACT

The integration of pre-trained deep language models, such as BERT, into retrieval and ranking pipelines has shown to provide large effectiveness gains over traditional bag-of-words models in the passage retrieval task. However, the best setup for integrating such deep language models is still unclear.

When BERT is used to re-rank passages (i.e., BERT re-ranker), previous work has empirically shown that, while in practice BERT re-ranker cannot act as initial retriever due to BERT's high query time costs, and thus a bag-of-words model such as BM25 is required. It is not necessary to interpolate BERT re-ranker and bag-of-words scores to generate the final ranking. In fact, the BERT re-ranker scores alone can be used by the re-ranker: the BERT re-ranker score appears to already capture the relevance signal provided by BM25.

In this paper, we further investigate the topic of interpolating BM25 and BERT-based rankers. Unlike previous work that considered the BERT re-ranker, however, here we consider BERT-based dense retrievers (RepBERT and ANCE). Dense retrievers encode queries and documents into low dimensional BERT-based embeddings. These methods overcome BERT's high computational costs at query time, and can thus be feasibly used in practice as whole-collection retrievers, rather than just as re-rankers.

Our novel empirical findings suggest that, unlike for BERT re-ranker, interpolation with BM25 is necessary for BERT-based dense retrievers to perform effectively; and the gains provided by the interpolation are significant. Further analysis reveals why this is so: dense retrievers are very effective at encoding strong relevance signals, but they fail in identifying weaker relevance signals – a task that the interpolation with BM25 is able to make up for.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; **Combination, fusion and federated search**; **Retrieval effectiveness**.

KEYWORDS

Dense Retrievers, BERT Ranking, Passage Retrieval, Neural IR

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

ACM ISBN 978-1-4503-8611-1/21/07…$15.00
https://doi.org/10.1145/3471158.3472233

ACM Reference Format:
Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. BERT-based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3471158.3472233

1 INTRODUCTION

Large, pre-trained, transformer-based deep language models such as BERT [6], T5 [19] and GPT [18], have been shown effective for text passage retrieval and ranking tasks [8, 14, 17, 26] when used as alternatives or in conjunction with conventional bag-of-words approaches such as BM25 [20]. Among these deep language models, BERT has so far received the lion share of attention from the research community. A common approach to integrate BERT within a retrieval pipeline is to use bag-of-words retriever such as BM25 (at times in combination with RM3 pseudo relevance feedback [1, 15]) for first stage retrieval, and then use BERT to re-rank BM25's top k passages (often, $k = 1,000$) [5, 9, 17]. (We refer to this approach as BERT re-ranker).

An implementation decision left to the search engine practitioner when implementing such BERT re-ranker is if it should use the output scores from BERT alone, or combine them with the original BM25 scores. The general BERT re-ranker allows both possibilities by defining the score $s(p)$ of a passage p as the interpolation between the two scores:

$$s(p) = \alpha \hat{s}_{BM25}(p) + (1 - \alpha) s_{BERT}(p) \quad (1)$$

where $\hat{s}_{BM25}(p)$ is the normalised BM25 score for passage p (see equation 16 by Lin et al. [14]), $s_{BERT}(p)$ is the BERT score for p, and the hyperparameter α controls the relative importance of BM25 and BERT scores.

We further note that the BERT re-ranker approach is general enough that could be used with emerging BERT-based representation learning methods that are alternative to BERT itself – commonly referred to as dense retrievers (DRs) [14]. Example of popular DRs are RepBERT [25] and ANCE [22], among others [7, 13, 16]. These alternative representations have been devised to address a limitation encountered when using BERT: The BERT inference step required at query time for each passage by the BERT re-ranker results in high query latency [14], making the method not scalable to full index retrieval, and still often impractical even for top k re-ranking (for large k values, e.g., $k = 1,000$). The common solution taken by DRs is to reduce the need for BERT inference at query time. This is often achieved by pre-computing BERT embeddings for passages at indexing time, although this means loosing the dependency between query and passage that is captured when

embeddings are created at query time. We note here that DRs can also be used as a second stage re-ranker on top of a bag-of-word model like BM25 [13, 16, 27].

Empirical evidence presented by Lin et al. [14] using the MS MARCO Passage Retrieval Task suggests that setting $\alpha = 1$, i.e., ignore the original BM25 scores when computing the final passage score, provides the highest level of effectiveness (as measured by MRR@10) among other choices for α. They go one step further, claiming that "exact term matching scores *[(BM25)], [...]*, do not provide any relevance signals that is not already captured by BERT". We further confirm these findings in Section 5 for the TREC DL 2019 and 2020, and for deep evaluation measures (MAP).

While interpolation with BM25 scores seems unnecessary for effective re-ranking with the BERT re-ranker, it is unclear if the same applies to novel BERT-based dense retrievers. DRs utilise the BERT representation to encode queries and passages in a low dimensional embedding [14]. These low dimensional passage embeddings can be constructed at indexing time, while the query embedding can be efficiently computed "on-the-fly" at query time [27]. This process makes it feasible to search the full passage collection, rather than performing a re-ranking task limited to a small number of top-k passages, like in the BERT re-ranker. Previous work usually considered these dense retrievers in isolation, i.e. without considering the interpolation with a bag-of-words model like BM25. However, the query latency offered by DRs is in the same order of that of BM25, and thus in practice the two could be executed in parallel once a query has been submitted to the search system, if their interpolation guarantees higher effectiveness than either method alone. We note that a recent DR model proposed by Gao et al. [10] does the linear interpolation with BM25 at the query inference time and has shown promising results; however, this model also requires interpolation with BM25 scores during training: a step that other DR models do not considered.

In this paper we provide a thorough empirical investigation of the importance of interpolating BERT and BM25 scores for those BERT-based DRs that do not consider BM25 scores during training and inference. We show that, unlike for the BERT re-ranker, this interpolation provides significant gains in effectiveness compared to using the BM25 or the DRs scores alone. We also show that DRs, when not interpolated with BM25, perform poorly with respect to deep evaluation measures, an aspect ignored in previous works, that have instead only focused on shallow measures (early rank positions). Finally, we provide evidence of why interpolation is required: our analysis in fact shows that BERT-based DRs are better than bag-of-words models (BM25) at modelling and detecting strong relevance signals. This is sufficient in providing good gains over BM25 for shallow measures. However, they fail to model the more complex, weaker relevance signals, for which instead BM25 does a good job: this results in DRs being outperformed by BM25 for deep measures. The interpolation of both methods is able to make up for each other's weaknesses.

2 RESEARCH QUESTIONS

To drive the empirical investigation put forward in this paper, we formulated the following research questions:

RQ1: Do **BERT-based dense retrievers** (DRs) encode the same relevance signal as BM25?

We investigate this by studying the optimal interpolation between dense retrievers scores and BM25 scores.

RQ2: Do the findings obtained for DRs for shallow evaluation measures **generalise to deep measures**?

We investigate this by considering deep relevance measures such as MAP, nDCG@1000 and recall@1,000.

RQ3: What level of effectiveness would be achievable if the **best interpolation setting** between DRs and BM25 scores **could be predicted** on a per query basis?

We investigate this by considering a query-by-query oracle system capable of providing the optimal interpolation parameter.

3 EXPERIMENTAL SETUP

To answer our research questions, we perform a thorough empirical investigation that considers different dense retrievers, passage ranking datasets, and both shallow and deep evaluation measures, representations. Next we provide the details of the experimental setup.

Datasets and Evaluation Measures. We use the following passage ranking datasets:

- MS MARCO Passage Dataset [2], which consists of 8.8M web page passages and more than 6K queries. We use MRR@10 and Recall@1000 as evaluation measure, in line with previous literature that has used MS MARCO. We note that, on this dataset, other measures will show similar findings and trends as MRR; this is because in this dataset a query is on average associated to only one relevant document. The MS MARCO dataset was used to derive the original claim by Lin et al. [14].
- TREC 2019 and 2020 Deep Learning Passage Retrieval Task (TREC DL) [3, 4], which relies on the same corpus used in MS MARCO but contains 43 and 54 queries, respectively. These datasets used deep judgement pools and graded relevance labels. As for evaluation measures, we report nDCG@10 as the shallow evaluation metric and nDCG@1000, MAP, Recall@1000 as deep evaluation metrics.

We use pairwise two-tails t-test with Bonferroni correction to measure statistical significant differences between retrieval runs. We consider significant the differences for which $p < 0.05$.

Rankers, Implementations, and Settings. For BM25, we rely on the implementation provided by Anserini [23], using the default parameters in the toolkit.

Aside from BM25, we investigate two BERT-based dense retriever methods, namely RepBERT [25] and ANCE [22]. We choose RepBERT and ANCE as representative BERT-based DRs because they provide nearly state-of-the-art effectiveness on the MS MARCO dataset and their implementations have been made publicly available. Our methodology can be easily adapted to other DR models such as the current state-of-the-art method RocketQA [7]; however, by the date of writing this paper, the implementation of RocketQA was not publicly available, and thus was not considered for this study.

For RepBERT, we use the implementation made available by the authors, but instead of the original self-implemented dot product

retrieval, we utilise the Faiss [11] toolkit to build the index and perform retrieval. For ANCE we use the scripts provided by the authors for both data pre-processing and model implementation.

In our experiments, we vary the interpolation parameter α from 0 to 1, with step of 0.1. We record results for each parameter value over the whole set of queries for a dataset. In addition, we also record the highest effectiveness achieved by any value of α on a per query basis: we use this to compute the effectiveness of an *"oracle"* system, i.e. a system that, for each query, could predict the value of α to set to obtain the highest effectiveness. We use this oracle to answer RQ3.

Along with the DR models we study in this paper, we also compare the results with CLEAR [10] as CLEAR also linearly interpolates the BM25 scores at the retrieval stage. In addition, the loss function used by CLEAR during training is designed for interpolating with BM25 to balance the bag-of-words signals provided by BM25 and the semantic matching signals provided by BERT. We note that although RepBERT and ANCE use BM25 results to sample hard negatives, they do not use any bag-of-words signals either during training or inference. Hence, in principle, RepBERT and ANCE may show a weaker ability to be interpolated with BM25. Nevertheless, in our experiments, we directly compare the effectiveness of RepBERT and ANCE with CLEAR when the first are interpolated with BM25 scores. Since the implementation and model of CLEAR was not publicly available at the time of writing this paper, we directly report the results from Table 1 of the original paper [10].

We make all implementations and analyses of our experiments publicly available at the GitHub repository https://github.com/ielab/InterpolateDR-ICTIR2021.

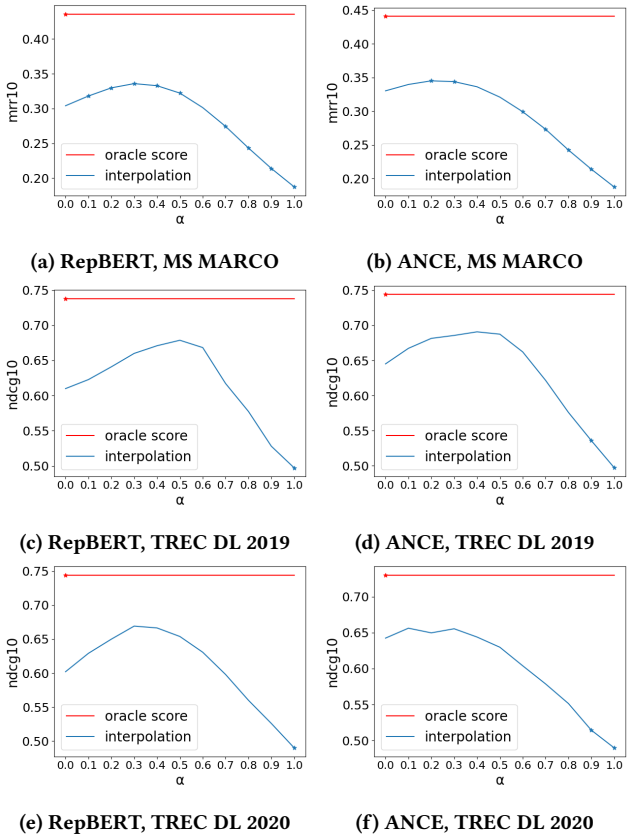
4 RESULTS

Next, we examine the results of our empirical investigation in light of the research questions put forward in Section 2.

4.1 RQ1: Do DRs encode the same relevance signal as BM25?

To answer RQ1, we consider the effectiveness of the BERT-based DRs by varying the BM25 interpolation parameters α. Following the interpolation setting for the BERT re-ranker reported by Lin et al. [14] which solely relies on shallow evaluation measures, we report MRR@10 for MS MARCO and nDCG@10 for TREC DL 2019 and 2020. In this experiment, we use both BM25 and DRs to retrieve the 2,000 highest scored passages and then perform the interpolation of the two lists of 2,000 results. This interpolation creates a combined result list with more than 2,000 passages; we then take the top 1,000 passages from the combined result list. A similar procedure is followed by Karpukhin et al. [12].

Results are reported in Figure 1 and Table 1. We first discuss the results obtained by DRs and BM25 alone, i.e. when no interpolation is used; these are obtained by setting $\alpha = 0$ for DRs, and $\alpha = 1$ for BM25. In this case, DRs methods obtain higher performance than BM25, at least in the shallow evaluation measures reported here. This result appears at first in like to that obtained by Lin et al. [14] for the BERT re-ranker, for the different task of passage re-ranking (recall that here instead we considered DRs retrieving from scratch, i.e. without relying on an initial BM25 ranking).



(a) RepBERT, MS MARCO (b) ANCE, MS MARCO

(c) RepBERT, TREC DL 2019 (d) ANCE, TREC DL 2019

(e) RepBERT, TREC DL 2020 (f) ANCE, TREC DL 2020

Figure 1: Results for the dense retrievers on MS MARCO, TREC DL 2020 and 2019 for varying values of interpolation parameter α (blue line). In red we display the effectiveness of the oracle system (see RQ3). $\star$ indicates statistical significant difference w.r.t. $\alpha = 0$.

Unlike the findings obtained for the BERT re-ranker, however, the highest effectiveness is not obtained by DRs alone: instead, higher effectiveness is achieved when DRs and BM25 scores are interpolated. For example, when $\alpha = 0.3$, the interpolation of DRs and BM25 scores achieves higher MRR@10 and nDCG@10 values than using DRs alone ($\alpha = 0$), regardless of the dataset or specific DRs used. We note that while gains obtained by interpolating DRs and BM25 are significant only for MS MARCO, the gains obtained for TREC DL 2019 and 2020 are still large, especially for RepBERT. Table 1 also reports the effectiveness of DRs, BM25 and their interpolation, with a fixed $\alpha = 0.5$, on all datasets for shallow evaluation measures; these results confirm the above findings.

The empirical results reported for RQ1 suggest that exact term matching methods such as BM25 can provide useful relevant signals to be added to DRs. This further highlights that the query and passage representations encoded by DRs fail to some extent to include exact term matching signal. This represents novel results not present in the literature, and in contrast to the findings obtained when the BERT re-ranker was considered. RepBERT and ANCE do not consider combining dense retrievers scores with those of

Table 1: Results for BM25, DRs and their interpolations ($\alpha = 0.5$ and oracle α) across all datasets with shallow evaluation metrics and their percentages of gains and losses of DRs over their corresponding non interpolated scores ($\alpha = 0$). Statistical significance differences are marked by †.

	MS MARCO	TREC DL 2019	TREC DL 2020
	MRR@10	nDCG@10	nDCG@10
BM25 ($\alpha = 1$)	0.1874	0.4973	0.4898
RepBERT ($\alpha = 0$)	0.3040	0.6100	0.6620
ANCE ($\alpha = 0$)	0.3302	0.6452	0.6424
RepBERT+BM25 ($\alpha = 0.5$)	0.3222(+6.0%)†	0.6787(+11.3%)	0.6539(-1.2%)
RepBERT+BM25 (oracle α)	0.4358(+43.4%)†	0.7380(+21.0%)†	0.7443(+12.4%)†
ANCE+BM25 ($\alpha = 0.5$)	0.3208(-2.8%)	0.6875(+6.6%)	0.6297(-2.0%)
ANCE+BM25 (oracle α)	0.4406(+33.4%)	†0.7441(+15.3%)	0.7301(+13.7%)
CLEAR [10]	0.338	0.699	-

Table 2: Results for BM25, DRs and their interpolations ($\alpha = 0.5$ and oracle α) across all datasets with deep evaluation metrics and their percentages of gains and losses of DRs over their corresponding non interpolated scores ($\alpha = 0$). Statistical significance differences are marked by †.

	MS MARCO	TREC DL 2019			TREC DL 2020		
	Recall@1000	nDCG@1000	MAP	Recall@1000	nDCG@1000	MAP	Recall@1000
BM25 ($\alpha = 1$)	0.8573	0.6001	0.3766	0.7384	0.5866	0.2870	0.7994
RepBERT ($\alpha = 0$)	0.9434	0.5986	0.3311	0.6689	0.5913	0.3704	0.7858
ANCE ($\alpha = 0$)	0.9582	0.6165	0.3611	0.6610	0.6301	0.4049	0.7733
RepBERT+BM25 ($\alpha = 0.5$)	0.9609(+1.9%)†	0.7172(+19.8%)†	0.4918(+48.5%)†	0.8128(+21.5%)†	0.6935(+17.2%)	0.4348(+17.4%)	0.8658(+10.2%)
RepBERT+BM25 (oracle α)	0.9706(+2.9%)†	0.7390(+23.5%)†	0.5254(+58.7%)†	0.8284(+23.8%)†	0.7272(+23.0%)†	0.4877(+31.7%)	0.8761(+11.5%)
ANCE+BM25 ($\alpha = 0.5$)	0.9697(+1.2%)†	0.7183(+16.5%)	0.4909(+35.9%)†	0.8136(+23.1%)†	0.6945(+10.2%)	0.4090(+1.0%)	0.8631(+11.6%)
ANCE+BM25 (oracle α)	0.9790(+2.2%)†	0.7429(+20.5%)†	0.5278(+46.2%)†	0.8335(+26.1%)†	0.7316(+16.1%)†	0.4836(+19.4%)	0.8768(+13.4%)
CLEAR [10]	0.969	-	0.511	0.812	-	-	-

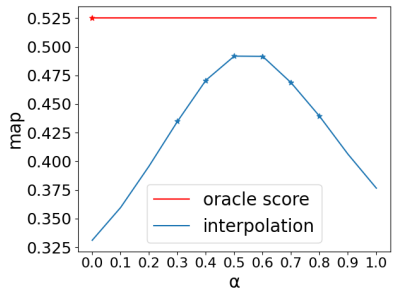
BM25: this combination, our results show, provides clear ranking improvements without hindering retrieval efficiency. In fact, both methods present similar query latency and can reasonably be run in parallel once the query is received.

4.2 RQ2: Interpolation results on deep evaluation measures

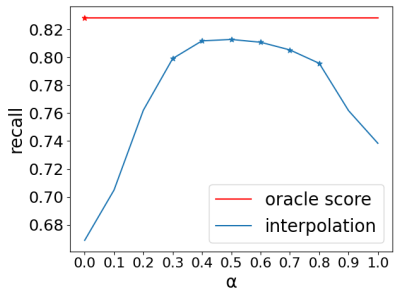
To answer RQ2, we consider the effectiveness of the BM25 and DRs with respect to deep evaluation measures such as MAP, nDCG@1000, and Recall@1000. Results are reported in Figure 2. Unlike findings for shallow evaluation measures, for which DRs always outperform BM25 (see RQ1), for deep measures we find that using DRs alone ($\alpha = 0$) does not always provide higher effectiveness than using BM25 alone ($\alpha = 1$). This is especially the case for Recall@1000 on TREC DL 2019 and 2020. While this observation does not hold true on MS MARCO for Recall@1000, we believe this to be due to an artefact of this specific dataset. MS MARCO in fact has a shallow judgment pools: on average, only one relevant passage per query, and typically with a high relevance label.

On the other hand, we find that the impact interpolating DRs scores and BM25 is much more significant on deep evaluation measures than on shallow measures. For example, when α is ≈ 0.5, deep effectiveness is much better than that of DRs alone ($\alpha = 0$) or of BM25 alone ($\alpha = 1$), see Figure 2. This is an important finding, as it largely differs from the previous results obtained for the BERT re-ranker. These results are also important because the likely usage of DRs is within a retrieval pipeline that considers further re-ranking steps after the use of DRs: thus high effectiveness on deep evaluation measures such as Recall@1000 is important for DRs.

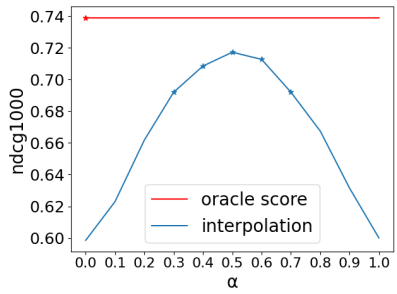
When comparing the results of RepBERT and ANCE with CLEAR (which does have a mechanism to interpolate the signals from BM25), it is clear that RepBERT and ANCE are much worse than CLEAR across all measures, when DRs are used without BM25 interpolation ($\alpha = 0$) in Table 1 and Table 2. However, after interpolating with BM25 ($\alpha = 0.5$), RepBERT and ANCE can achieve the same level of performance as CLEAR for both shallow and deep evaluation metrics: for Recall@1000 RepBERT and ANCE can even surpass CLEAR. While RepBERT and ANCE are not specifically designed to exploit BM25 signals as CLEAR instead does, they do obtain major benefits from the interpolation, suggesting that interpolating with BM25 is important even for DRs that are not specially designed to do so.



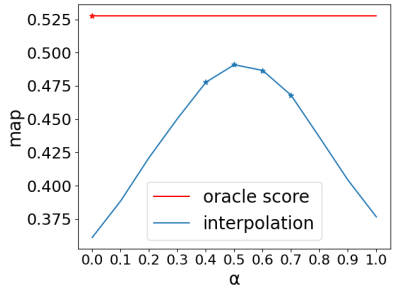
(a) RepBERT, TREC DL 2019, MAP



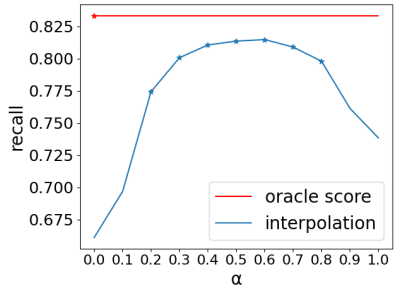
(b) RepBERT, TREC DL 2019, Recall@1000



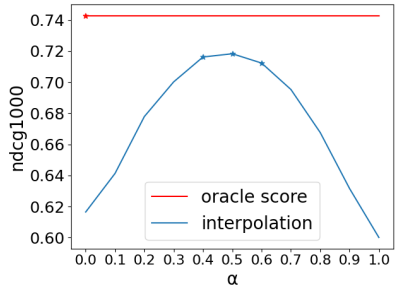
(c) RepBERT, TREC DL 2019, nDCG@1000



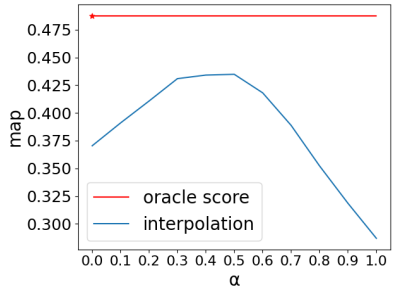
(d) ANCE, TREC DL 2019, MAP



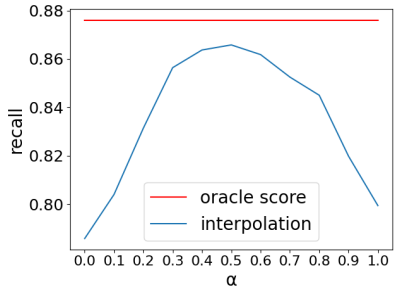
(e) ANCE, TREC DL 2019, Recall@1000



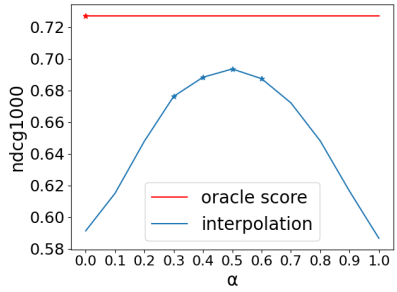
(f) ANCE, TREC DL 2019, nDCG@1000



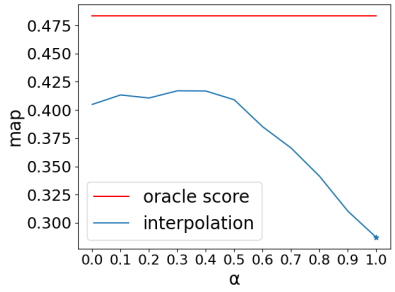
(g) RepBERT, TREC DL 2020, MAP



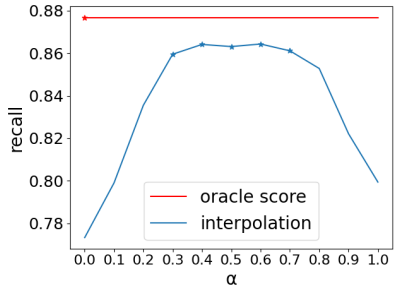
(h) RepBERT, TREC DL 2020, Recall@1000



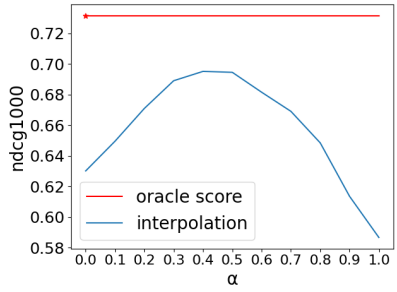
(i) RepBERT, TREC DL 2020, nDCG@1000



(j) ANCE, TREC DL 2020, MAP



(k) ANCE, TREC DL 2020, Recall@1000



(l) ANCE, TREC DL 2020, nDCG@1000

Figure 2: Results for the dense retrievers on TREC DL 2019 and 2020 for varying values of interpolation parameter α (blue line). In red we display the effectiveness of the oracle system. $\star$ indicates statistical significant difference w.r.t. $\alpha = 0$.

4.3 RQ3: Upper bound on interpolation

To answer RQ3, we consider the effectiveness of the oracle system – a system that would be able to determine a priori the most effective value of α for every query. This provides an indication of the upper bound effectiveness that could be achieved by DRs and their interpolation with BM25.

The oracle results for the settings analysed thus far have been reported in the figures for the previous research questions: they are

Table 3: The distributions of relevant passages in the top 100 results with respect to level of relevance for TREC DL 2019 and 2020. Level 1 means marginally relevant, 2 means relevant and 3 means highly relevant.

	TREC DL 2019			TREC DL 2020		
	$rel = 1$	$rel = 2$	$rel = 3$	$rel = 1$	$rel = 2$	$rel = 3$
BM25	44.0%	44.0%	48.1%	38.6%	46.3%	58.1%
RepBERT	31.1%	47.8%	58.4%	31.2%	49.3%	63.0%
ANCE	35.0%	50.0%	57.5%	35.7%	53.0%	67.8%

marked with a red line. In addition, both Table 1 and Table 2 provides a summary comparison between BM25, which is equivalent to the setting $\alpha = 1$, the BERT-based DRs which only considers BERT scores ($\alpha = 0$), the DRs ranker with $\alpha = 0.5$ which we find to be robust on deeper ranking metrics across all datasets, and the oracle system. This comparison is reported across DRs (RepBERT, and ANCE) and for all datasets and evaluation measures.

From Table 1 and Table 2 we observe, somewhat unsurprisingly, that the oracle system delivers large, significant gains over the other settings: this is done by identifying for each query the most effective value of the interpolation α. We also remark that, again, tuning $\alpha = 0.5$, rather than setting it to 0, provides improvements for all datasets with exceptions on shallow evaluation metrics.

Also, our results, when compared to CLEAR, show that CLEAR does not combine these two signals in the most optimal way, i.e. RepBERT/ANCE + BM25 is better than CLEAR when using oracle α. This suggests that CLEAR may also have large margins of improvements.

5 DISCUSSION

5.1 Why interpolation matters?

Unlike for the BERT re-ranker, the empirical results obtained for DRs have shown that the best effectiveness is achieved when their score is interpolated with that of BM25. This suggests that the DRs and BM25 encode different relevance signals. In particular, DRs exhibit lower recall at deep ranks (as indicated by MAP and Recall@1000), despite high ranking effectiveness in the early rank positions (as indicated by nDCG@10).

To understand why this is the case, we investigate the methods ability to retrieve passages cross different levels of relevance. We report this analysis for TREC DL 2019 and 2020, while this analysis cannot be performed for MS MARCO (relevance not graded, and on average there is one relevant document per query only). Table 3 reports the results of this analysis, where the percentages are computed with respect to the ratio between the number of relevant passages with a specific grade retrieved by the method and the number of relevant passages with that specific grade present in the collection. For example, the value of 48.1% for level 3 obtained by BM25 means that BM25 retrieved 48.1% of all highly relevant documents for the collection in the top 100 rank positions.

According to the results of Table 3, BM25 is fairly unbiased in retrieving passages across the different levels of relevance, when TREC DL 2019 is considered. In fact, for this collection, BM25 retrieves a similar percentage of the passages of relevance level 3 (highly relevant) as it does for level 2 (relevant), and for level 1 passages (marginally relevant). Conversely, the studied DRs exhibit a bias towards preferring relevant and highly relevant passages, e.g., retrieving 57.5% or more of the highly relevant passages in the top 100 rank positions vs. 31% - 35% for the marginally relevant passages. These results are similar in TREC DL 2020: although the distribution for BM25 is more imbalanced towards more relevant passages, compared to that obtained on TREC DL 2019, the bias measured for BM25 in this collection is less than that expressed by DRs. Similar results are obtained when considering other rank cut-offs (e.g., top 10 ranks, or top 1,000 ranks).

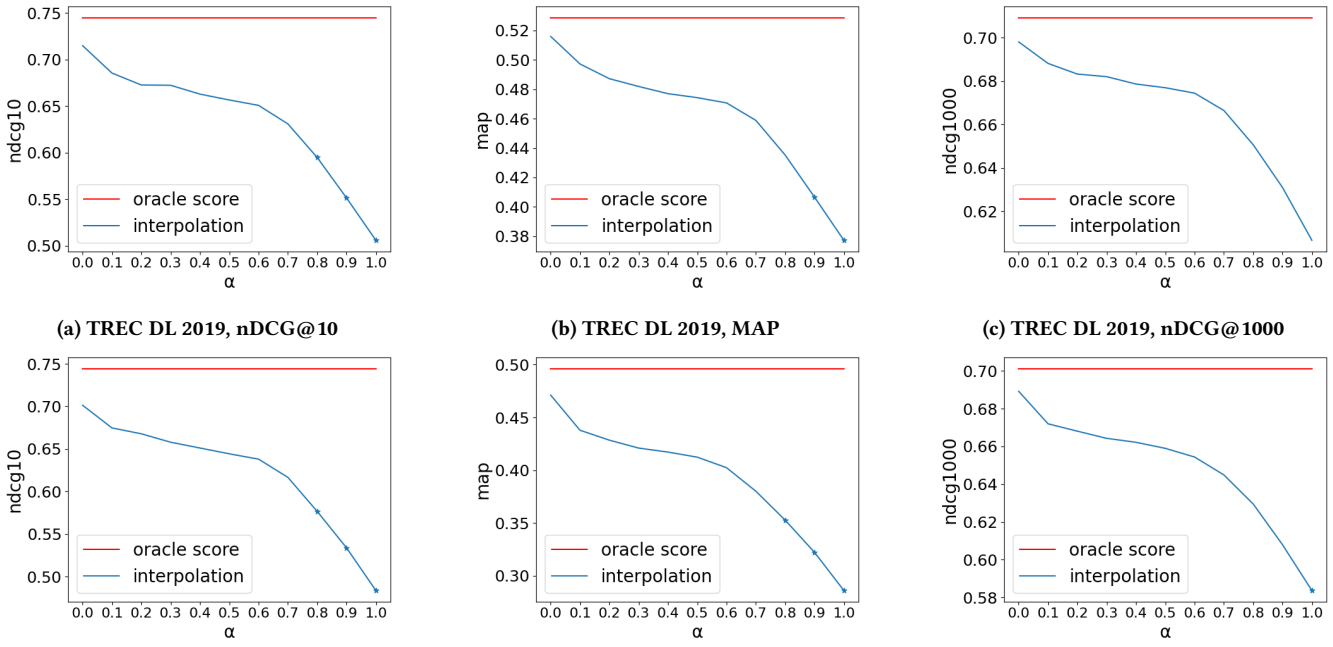
We have two hypothesises to why RepBERT and ANCE have a bias towards highly relevant passages, performing badly for retrieving weak relevant passages (and thus overall delivering a lower recall):

(1) This may due to the bias that exists in the dataset used to train the dense retrievers. Both RepBERT and ANCE, in fact, are trained with the MS MARCO passage ranking dataset, which only provides one relevant passage per training query on average – and this one passage is highly relevant. However, there are usually many passages that are relevant, to different extent and with differing aspects, for the same query. Hence, the dense retrievers trained with only one relevant passage for each query may not form a good enough representation of relevance to be able to better distinguish those passages with a weak relevance signal.
(2) We note that both RepBERT and ANCE use BM25 to sample hard negatives passages. More specifically, they randomly sample passages from the top passages retrieved by BM25 and then treat these passages as irrelevant to train the dense retriever. We think hard negative sampling strategy may be dangerous as there may be lots of false negatives retrieved by BM25 (i.e. actual relevant passages). In support to this hypothesis are recent studies that have shown that de-noising the hard negatives during the training of dense retrievers is important [7] and BM25 negative sampling often hurts the dense retrievers effectiveness in terms of recall [24].

In summary, this analysis reveals that DRs (especially trained with the BM25 hard negative sampling strategy) are highly effective in encoding strong relevance signals – indeed, they do a better job than BM25 at doing so. However, they are not as good when it comes to modelling weaker relevance signals, and in turns BM25 outperforms DRs in this task. It is the interpolation of both methods, however, that is able to make up for each other's weaknesses, as shown in the results of Section 4 – and this interpolation is a relatively a simple solution compared to more sophisticated techniques [10].

5.2 BERT re-ranker, other datasets and deep measures

Our experiments tested BERT-based dense retrievers and their interpolation with BM25 on multiple datasets and by considering both shallow and deep evaluation measures. The interpolation between BM25 and the BERT re-ranker, however, was only empirically validated on the MS MARCO dataset and interpolation was shown



(a) TREC DL 2019, nDCG@10 (b) TREC DL 2019, MAP (c) TREC DL 2019, nDCG@1000

(d) TREC DL 2020, nDCG@10 (e) TREC DL 2020, MAP (f) TREC DL 2020, nDCG@1000

Figure 3: Results for the BERT re-ranker on TREC DL 2019 and 2020 for varying values of interpolation parameter α (blue line). In red we display the effectiveness of the oracle system. $\star$ indicates statistical significant difference w.r.t. $\alpha = 0$.

unnecessary for shallow evaluation measure but was untested for deeper measures. Next, we complement these experiments, by considering the effectiveness of the BERT re-ranker by varying the interpolation parameter α across the passage retrieval datasets considered in our main experiments for dense rankers and reporting also deep evaluation measures.

To implement the BERT re-ranker, we extend the Anserini toolkit using the Huggingface implementation [21] of mono-BERT large provided by Lin et al. [17]; this model is fine-tuned on the MS MARCO passage dataset. We use the top 1,000 results retrieved by BM25 for re-ranking.

For all considered datasets (TREC DL 2019 and 2020), we obtained similar findings to those of Lin et al. [14], i.e. that the best value is obtained when only BERT and not BM25 is used for re-ranking, see Figure 3. This is regardless of the evaluation measure used: note for example that the trends for α observed in Figure 3 for nDCG@10 are similar to those observed for MAP and nDCG@1000.

We further note that the ability to set α to an optimal value on a per query basis (oracle system, red lines in Figure 3) does return further improvements in effectiveness; however these improvements are lower than those attainable by the oracle system when DRs are used.

Finally, we note that DRs have been proposed as feasible full-index or re-ranking alternatives to the BERT re-ranker, which often cannot be run at runtime, unless highly constrained in the amount of passages considered for re-ranking. Previous empirical results have suggested that DRs trade runtime improvements for a decreased ranking quality, compared to the BERT re-ranker. However, this is not often the case when DRs scores are interpolated with the BM25 scores: this is true in particular when deep evaluation measures are considered in place of shallow ones.

6 CONCLUSION

The use of BERT within retrieval pipelines has shown to significantly improve the effectiveness of traditional bag-of-words approaches. Previous work has found that a simple ranker that uses BM25 for first stage retrieval and re-ranks its top results using solely relevance signal from BERT is more effective than combining both neural (BERT) and bag-of-words (BM25) relevance signals. From this, it was claimed that BERT incorporates (and further extends) the relevance signal provided by bag-of-words models. In this paper we investigate whether this finding and associated claim extend to alternative BERT-based representations, called dense retrievers.

Our empirical investigation finds these earlier results obtained for the BERT-reranker not to hold when BERT-based DRs are considered: the interpolation between BM25 scores and DRs is important, as it provides higher effectiveness than each of the methods alone, across both shallow and deep evaluation measures, and across a range of datasets. This is a novel and important result because, in practice, the BERT re-ranker has displayed high query latency [14] (in the order of thousands of milliseconds), which make it challenging for use in practical search engine applications. DRs like RepBERT and ANCE have been designed to allow more feasible runtime (in the order of tens, for re-ranking, to hundred milliseconds, for full-index retrieval, like that considered in this paper). Thus, DRs are more likely to be used in real applications

than the BERT re-ranker: yet they behave differently from the BERT re-ranker itself with respect to the interpolation parameter α. We also note that because of their feasible runtime, DRs are likely be used at the early steps of cascade ranking architectures – thus a high effectiveness on deep evaluation measures, like that provided when DRs are interpolated with BM25 scores, is important, as it would influence the downstream effectiveness obtained by more expensive re-rankers, as measured by shallow measures.

Finally, our empirical investigation has also studied the optimal value of α on a per query basis, for both BERT-based DRs and the BERT re-ranker. The results suggest that substantial gains are obtainable if adequate methods to predict the optimal value of α given a query were available. This prediction task is a possible direction we aim to pursue in future work.

7 ACKNOWLEDGEMENT

Dr Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579). This research is partially funded by the Grain Research and Development Corporation project AgAsk (UOQ2003-009RTX). The authors are thankful to Mr Hang Li (ielab, UQ) for help with implementation of the dense retrievers methods used in this work.

REFERENCES

[1] Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. *Computer Science Department Faculty Publication Series* (2004), 189.

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 Deep Learning Track. (2020).

[4] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2021. Overview of the TREC 2020 Deep Learning Track. (2021).

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

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-Training of Deep Bidirectional Transformers For Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Yingqi Qu Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. *arXiv preprint arXiv:2010.08191* (2020).

[8] Cicero dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Beyond [CLS] through Ranking by Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1722–1727.

[9] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Rethink training of BERT rerankers in multi-stage retrieval pipeline. In *Proceedings of the 43rd European Conference On Information Retrieval (ECIR)*.

[10] Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement Lexical Retrieval Model with Semantic Residual Embeddings.. In *ECIR (1)*. 146–160.

[11] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* (2019).

[12] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[13] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 39–48.

[14] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond. *arXiv preprint arXiv:2010.06467* (2020).

[15] Xiaoyong Liu and W Bruce Croft. 2002. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*. 375–382.

[16] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Expansion via prediction of importance with contextualization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1573–1576.

[17] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[18] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[20] Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

[21] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.

[22] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations*.

[23] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *Journal of Data and Information Quality (JDIQ)* 10, 4 (2018), 1–20.

[24] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[25] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized Text Embeddings for First-Stage Retrieval. *arXiv preprint arXiv:2006.15498* (2020).

[26] Shengyao Zhuang, Hang Li, and Guido Zuccon. 2021. Deep Query Likelihood Model for Information Retrieval. In *Proceedings of the 43rd European Conference On Information Retrieval (ECIR)*.

[27] Shengyao Zhuang and Guido Zuccon. 2021. TILDE: Term Independent Likelihood moDEl for Passage Re-ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.