University of Amsterdam at TREC 2021: Deep Learning Track

Jaap Kamps David Rau

University of Amsterdam

Abstract

This paper documents the University of Amsterdam's participation in the TREC 2021 Deep Learning Track. In addition to providing labeled training data at scale, the other major contribution of the TREC DL track is to avoid the pool-bias exhibited in all the earlier adhoc search test collections created through the pooling only runs from traditional sparse retrieval systems. However, even in the TREC deep learning track, we have shallow pools, and runs with varying but high fractions of unjudged documents. This prompts a deeper analysis of pool coverage over ranks for both a representative traditional approach (i.e., BM25) and a representative neural approach (i.e., the BERT cross-encoder for the passage retrieval task). Our main conclusions are the following. First, we submitted a neural run that specifically looks beyond those documents easily found by traditional models, highlighting the potential of neural models to address recall-aspects in addition to the precision aspects prioritized in the TREC Deep Learning Track up to now. Second, we observe high fractions of unjudged documents after the initial ranks for both the 2020 and 2021 data, which may hinder the evaluation of recall-oriented aspects and reusability of the judgments for runs not contributing to the pooling. Third, we observe a gradual decline of the fraction of relevant over judged documents for 2020, which is a positive sign against pooling bias, but almost no decrease for 2021. Our general conclusion is that coverage below the guaranteed pooling horizon is far from complete and that analysis of recall aspects must be done with care, but that there is great potential to study these in future editions of the track.

1 Introduction

This paper documents our participation in the TREC 2021 Deep Learning Track. The Deep Learning Track started at TREC 2019 [Craswell et al., 2020, 2021], and is in it's third year. The TREC Deep Learning Track is providing very large training data for its task, which is a necessity for training deep neural networks. The track consists of four tasks in total; two tasks for two collections. Full-Ranking and Re-Ranking tasks applied in collections of documents and passages. The Re-Ranking task focuses on ranking the top 1,000 candidates from the collection for each query. We focused on the collection of passages and performed experiments for both tasks. TREC provides training, evaluation and test queries, the collection of passages and training triplets that consist of a query, a couple of candidates and an indicator that defines the most relevant passage among the two candidates.

This paper is structured in the following way. Our simple experiment is described in Section 2 and the results of these experiments in Section 3. Finally, we end in Section 4 with a discussion of our main findings.

2 Experimental Design

In this section we detail our simple pooling experiments.

In addition to providing labeled training data at scale, the other major contribution of the TREC DL track is to avoid the pool-bias exhibited in all the earlier adhoc search test collections created through the pooling only runs from traditional sparse retrieval systems. However, even in the TREC deep learning track, we have shallow pools and runs with varying but high fractions of unjudged documents.

Although all official submissions to the TREC deep learning track contribute to the pool, only the top 10 of the run is guaranteed to be judged by the assessors, and the remaining pool is determined incrementally by an active learning approach based on the already judged passages or documents up to this point. This leads to varying, and sometimes high, fractions of unjudged documents over official submissions, and even high fractions of unjudged documents in the top results of post-submission experiments with runs not contributing to the pool. So far this is based on anecdotal evidence from the analysis of recall aspects of various models [e.g., our own earlier work, Rau et al., 2020].

This prompts a deeper analysis of pool coverage over ranks for both a representative traditional approach (i.e., BM25) and a representative neural approach (i.e., the BERT cross-encoder for the passage retrieval task).

BM25 BM25 is the prototypical traditional system. As in 2021 the reranking task only provides a top 100 rerank set based on BM25, we reconstruct the BM25 run using the Anserini setup. Specifically, we use no stemming, standard stopword removal, and don't optimize the BM25 parameters. This results in marginally lower scores than an optimized setup, and marginally lower scores than a feedback variant with RM3 [e.g., Craswell et al., 2021].

CE The BERT cross-encoder is a prototypical application of large pre-trained transformers, in particular BERT, to the passage ranking task by directly encoding a query+passage and training class label that informs the ranking task. Specifically, we use the pretrained BERT and train the CLS token in an additional layer using the MS-Marco training data. This model has shown significant and clear improvements over traditional systems in the track [Craswell et al., 2020, 2021].

As our main interest is to do analysis that may generalize to other settings, we strictly prefer to work with a setup that is as clean and simple as possible, without optimizing the settings to specific setup and conditions of the TREC track.

Recall that only the top 10 is guaranteed to be pooled, so in order see beyond this shallow pooling horizon, we are particularly interested in what happens at a lower rank (say, beyond the top 100). We specifically manipulate a submission to contain only results that are outside the top 100 of the traditional BM25 run (corresponding to the rerank task in 2021). In this way, we hope to increase pool diversity, and also obtain a comparable sample of judged documents from lower in the ranking: what is the quality of the neural ranker beyond the initial precision?

3 Experimental Results

This section contains the main results of our experiments, focusing on the analysis of pool coverage over standard sparse retrieval systems and standard neural rankers.

Table 1: Performance on the NIST judgments of the TREC Deep Learning Task 2020

Ranker	NDCG@10	MAP	MRR	P@10
BM25	49.59	27.47	67.06	37.96
Cross-encoder	68.95	45.96	79.88	50.93

Table 2: Performance on the NIST judgments of the TREC Deep Learning Task 2021

Ranker	NDCG@10	MAP	MRR	P@10
BM25 (official provided)	31.60	21.22	84.53	35.47
BM25 (own)	30.58	18.43	82.99	33.96
Cross-encoder (official top-100)	59.98	18.83	75.23	50.19
Cross-encoder (w/o official top-100)	51.04	12.17	65.66	43.21

3.1 Effectiveness

As our main goal is the analysis of pool coverage, we opt to study two runs in detail. First, a standard BM25 runs as used as an oracle for the rerank sets in the passage retrieval task in 2020 (top 1,000 passages) and 2021 (top 100 passages). As the 2021 rerank set contains fewer documents, we use similar top 1,000 runs in order to compare both years. Second, a standard BERT / cross-encoder model, that is a clean and simple model that incorporates the effect of using large language models (in particular pretrained BERT) in a passage retrieval setting.

Tables 1 and 2 show the effectiveness of these two models. In this task setting, the cross-encoder clearly outperforms the traditional BM25 system on the main measure NDCG@10, as was observed throughout the years [Craswell et al., 2020, 2021]. While the BM25 run has the highest initial precision (with an MRR of 83-85% versus 75% for the neural model), the neural model clearly outperforms on NDCG@10, also reflecting the superiority of dense models to correctly rank passages with the highest "Perfect" assessment. It is worthy to note that this years cross-encoder scores are significantly lower than last year. This can potentially be explained by the significant increase of the collection containing orders of magnitude more documents this year, and the use of a smaller "rerank" set of 100 passages leading to more diversity in the submissions.

Our only official submission was a cross-encoder run which left out any document already in the top 100 rerank set of the passage retrieval task. As expected removing the top-100 documents (see Table 1 `w/o official top-100`) re-rank set decreases the AP scores significantly, as we're focusing exclusive on the harder to find relevant documents (ignoring all that BM25 can pick up with ease). Looking at the top of the manipulated run, we see only a small decrease in initial precision (NDCG@10 decreases from 60% to 51% and P@10 from 50% to 43%).

This manipulated run intended to shed some new light on what's happening in the neural ranking beyond the shallow pooling depth used in the TREC Deep Learning track, motivated by earlier observations that relatively high fractions of unjudged documents remain, even in the rerank setup of the passage retrieval task. While this filtered run leaving out the "easy" relevant documents takes an obvious hit in performance, and isn't a contender for the top of the table in terms of performance, it still obtains very reasonable precision scores. This highlights the potential of neural models to address recall-aspects in addition to the precision aspects prioritized in the TREC Deep Learning Track setup up to now.

Table 3: Judged passages across ranks (TREC Deep Learning Track 2020)

		1	5	10	50	100	500	1,000
BM25	Relevant (%)	77.78	65.19	56.67	31.26	21.52	6.91	3.92
	Non-relevant (%)	22.22	34.44	40.00	36.22	27.06	10.13	6.07
	Unjudged (%)	0.00	0.37	3.33	32.52	51.43	82.96	90.01
	Rel./Judged (%)	77.78	65.43	58.62	46.32	44.30	40.58	39.23
CE	Relevant (%)	87.04	80.37	71.67	42.93	29.04	7.63	3.90
	Non-relevant (%)	12.96	18.89	25.56	30.41	25.65	10.80	6.45
	Unjudged (%)	0.00	0.74	2.78	25.85	43.98	74.73	80.67
	Rel./Judged (%)	87.04	80.97	73.71	58.54	53.10	41.41	37.65

3.2 Analysis

We focus our analysis on two prototypical runs: BM25 to represent a traditional sparse retrieval model, and the BERT cross-encoder to represent an effective neural dense retrieval model relying on large language models such as BERT.

Table 3 shows a breakdown of judged passages over the ranks for the 2020 qrels. Here, relevant means any degree of relevance (including 'Related'), whereas the evaluation scores use only the 'Highly Relevant' and 'Perfect' judgments. In 2020, the rerank task was based on the BM25 top 1,000 passages provided, with full-rank systems obtaining very similar recall as rerank submissions [Craswell et al., 2021]. In terms of pooling: the organizers pooled the top 10 of all submitted runs, and used active learning (HiCal BMI) for incrementally expanding the pool based on the relevant and non-relevant passages up to this point. A total of 54 topics with at least three known relevant and a ratio of relevant over judged below 40 percent form the 2020 test collection.

BM25 and cross-encoder runs were submitted by some teams in 2020, as runs `p_bm25` and `nlm-bert-rr` respectively in [Craswell et al., 2021], hence contributed to the pool but with only their top 10's being guaranteed to be included. Considering BM25 we see a very low percentage of unjudged documents within the first 10 ranks (3.3%@10). However, when looking at lower ranks the percentage of unjudged increases dramatically from 46.32% @50 up to 90.01% @1,000. For CE we observe a similar trend: Up to rank 10 very few documents are unjudged (2.78%), whereas later in the 25.58% @50 up to 80.67% @1,000 are unjudged. Those numbers show that the judgments of the neural model exhibit higher coverage compared to the traditional model in 2020.

With so many unjudged documents, are we seriously underestimating the performance? We also look at the fraction of Relevant to Judged documents for both types of runs, and observe that fraction goes down over ranks suggesting that the pool is not unfavorably biased against the runs. We also observe that the neural model is superior compared to the traditional model, although at the end of the ranking both sets converge due to both reranking the same top 1,000 of passages.

Table 4 replicates this for the 2021 judgments. Recall that in 2021, the rerank task was based on only the top 100 based on a BM-25 run, likely leading to far greater run diversity. To do a comparable analysis, we reconstructed a top 1,000 rerank set (see Table 2 `BM25 (own)`). In terms of pooling, again the organizers automatically pooled the top 10 of all submitted runs, and use of active learning (CAL) for incrementally expanding the pool based on the relevant and non-relevant passages up to this point. A total of 53 topics with at least five known relevant documents, combined with sufficient judgements and low enough generality, form the 2021 test collection. Unlike 2020, the organizers already issue a warning in 2021 that even some official submissions, with guaranteed

Table 4: Judged passages across ranks (TREC Deep Learning Track 2021)

		1	5	10	50	100	500	1,000
BM25	Relevant (%)	75.47	69.81	66.04	40.34	29.81	10.94	6.57
	Non-relevant (%)	22.64	26.04	24.91	18.68	13.94	4.90	2.95
	Unjudged (%)	1.89	4.15	9.06	40.98	56.25	84.16	90.49
	Rel./Judged (%)	76.92	72.83	72.61	68.35	68.13	69.08	69.03
CE	Relevant (%)	75.47	73.58	70.57	34.49	22.57	6.61	3.43
	Non-relevant (%)	24.53	26.42	29.43	11.92	7.66	2.41	1.33
	Unjudged (%)	0.00	0.00	0.00	53.58	69.77	90.98	86.48
	Rel./Judged (%)	75.47	73.58	70.57	74.31	74.66	73.26	72.10

top 10 pools, have high fractions of un-judged documents below this shallow pooling cut-off.

Judging by the pool coverage, BM25 and cross-encoder runs were submitted by some teams in 2021. We only submitted a manipulated version of the CE run that left out documents from the official top 100 BM25 rerank set, in order to increase pool diversity and to aid analysis of the effectiveness of the neural ranker beyond the pooling horizon.

Considering BM25 we see, similarly to 2020, a very low percentage of unjudged documents within the first 10 ranks (9.06% @10), with the percentage of unjudged increasing from 40.98% @50 to 90.49% @1,000. For CE we observe a similar trend: Up to rank 10 the percentage of unjudged is zero as the top-10 contributed to the pool. However, later in the ranking 53.58% @50 up to 86.48% @1,000 are unjudged. So similar to 2020, we observed very high fractions of unjudged documents deeper in the rankings. We also look again at the fraction of relevant over judged documents for both runs. Unlike 2020, we observe in 2021 an almost non-decreasing fraction of relevant documents over the ranks, for both BM25 and the neural model. The BM25 fractions marginally drop from 77% to 69%, and the neural model from 75% to 72%. This could be a signal that the pools are just too shallow to cover a representative (and unbiased) sample of the relevant documents.

We also look at the pool coverage of our manipulated neural run looking beyond the rerank set's BM25 top 100 (not displayed in the Table). When comparing the runs of 2020 and 2021, we observe that our manipulated run contains twofold more unjudged documents than the regular neural model in Table 4, suggesting that documents that are brought up from lower ranks are not covered well by the current form of pooling.

The pool incompleteness is a necessary consequence of finite judging budgets, and mostly a call to caution when interpreting the results of the evaluation. Our analysis signals that pool coverage below the guaranteed pooling horizon is far from complete, and that analysis of recall aspects and runs not directly contributing to the pools, must be done with care.

4 Conclusions

This paper documented our participation in the TREC 2021 Deep Learning Track. We conducted a simple experiment in quantifying pool coverage. Our main conclusions are the following. First, we submitted a neural run that specifically looks beyond those documents easily found by traditional models, highlighting the potential of neural models to address recall-aspects in addition to the precision aspects prioritized in the TREC Deep Learning Track up to now. Second, we observe

high fractions of unjudged documents after the initial ranks for both the 2020 and 2021 data, which may hinder the evaluation of recall-oriented aspects and reusability of the judgments for runs not contributing to the pooling. Third, we observe a gradual decline of the fraction of relevant over judged documents for 2020, which is a positive sign against pooling bias, but almost no decrease for 2021. Our general conclusion is that coverage below the guaranteed pooling horizon is far from complete and that analysis of recall aspects must be done with care, but that there is great potential to study these in future editions of the track. At the time of writing, the systems contributing to the pools have not been released, and more advanced reusability analysis will be added in a later stage of this paper.

Acknowledgments

We thank the track organizers for their amazing service and effort in making realistic benchmarks for neural ranking available. This research is funded in part by the Netherlands Organization for Scientific Research (NWO STW # 17752; NWO CI # CISC.CC.016), Facebook Research (Computationally Efficient NLP grant), and the Innovation Exchange Amsterdam (POC grant). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and E. M. Voorhees. Overview of the TREC 2019 deep learning track. In *TREC 2019: Proceedings of the Twenty-Eighth Text REtrieval Conference*. NIST Special Publication 1250, 2020.

N. Craswell, B. Mitra, E. Yilmaz, and D. Campos. Overview of the TREC 2020 deep learning track. In *TREC 2020: Proceedings of the Twenty-Ninth Text REtrieval Conference*. NIST Special Publication, 2021.

D. Rau, N. Kondylidis, and J. Kamps. University of Amsterdam at TREC 2020: Deep learning track. In *The Twenty-Ninth Text REtrieval Conference Notebook (TREC 2020)*. National Institute for Standards and Technology, 2020.