Evaluation of Models for Ranking of Long Documents (online appendix)

Anonymous Author(s)*‡‡‡

1 SENSITIVITY ANALYSIS/ABLATIONS

1.1 Truncation threshold

Table 1: Effect of Document Truncation on Accuracy (Robust04 and PARADE Attn)

# of BERT chunks	1	2	3	4	5	6
MAP	0.278	0.302	0.314	0.317	0.320	0.317
NDCG@20	0.478	0.501	0.509	0.509	0.511	0.510

In our study, we truncate all the documents to have at most 1431 tokens, which correspond to three chunks each containing 477 document tokens, up to 32 query tokens, and three special tokens (`[CLS]` and two `[SEP]`). In preliminary experiments, we were not able to achieve any gains on MS MARCO by considering six chunks. Here, we asses the effect more systematically for Robust04 and *PARADE-Attn* model. As in the main experiments, this model is first trained on MS MARCO and then fine-tuned on Robust04. While doing so, we use the same truncation threshold during both training and testing steps. As we can see from Table 1, increasing the maximum input length beyond 1431 tokens (three chunks), has only marginal effect on accuracy (less than 2% gain in MAP and only 0.3% gain in NDCG@10).

1.2 A Surprising Effectiveness of *FirstP* baselines: Is Our Data Biased?

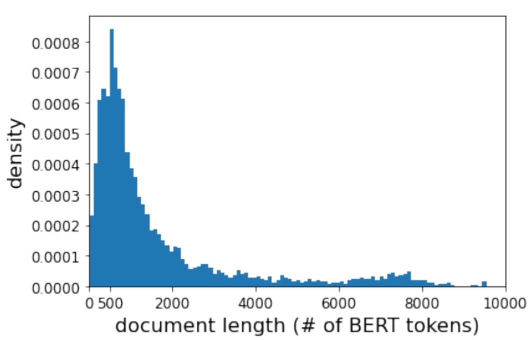
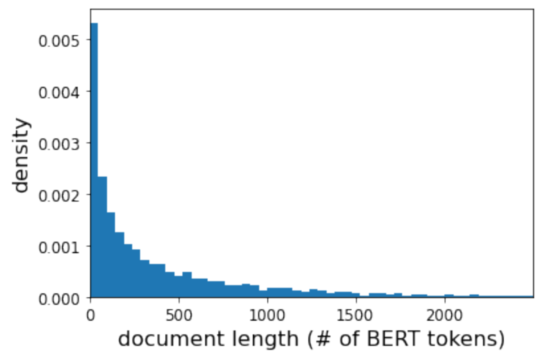
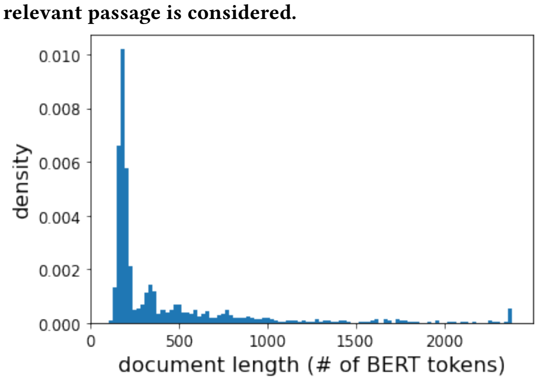


Figure 1: Distribution of relevant document lengths (in BERT tokens) on MS MARCO (v1) development set (doc. length is capped at 10K tokens)

As we could see from main experimental result in Table 3, the full-document models only marginally outperform the respective *FirstP* baselines: The BEST performing *LongP* models outstripped respective *FirstP* baselines by only 4-6%. As we show in § 1.1, this is not due to truncation of documents: Doubling the maximum input length does not further increase the accuracy of the *PARADE Attn* model. To shed light on this phenomenon, we plot the distribution of relevant document lengths (Fig. 1) and compare it with the positions of the first relevant passage in a document (Fig. 2). Both are measured in the number of BERT tokens.



(a) Development set (estimated positions). Only the first relevant passage is considered.



(b) TREC DL 2019 query set and crowd-source positions (FIRA data: Only the first relevant passage is considered. All positive relevance grades are included)

Figure 2: Distribution of *ending* positions (in # of BERT tokens) of relevant passages inside documents on two MS MARCO v1 query sets.

As a reminder, document-level relevance labels in MS MARCO v1 are created by transferring passage-level relevance to original documents from which passages were extracted. However, this mapping is not provided and we attempted to recreate it automatically. Because passage and document collections were gathered at different times, document texts diverged from the their original versions and, thus, exact matching of passages to documents is generally impossible. In particular, Hofstätter et al. [2] were able to match only 32% of the passages. We were able to obtain a much more representative matching statistics (for about 85% of the passages) using approximate matching.

Specifically, we used a combination of finding a longest common substring (threshold 0.8) as a primary matching option and a longest common subsequence (threshold 0.7) as a fallback option.[1] Please note that finding in-document passage matches is prohibited for the purpose of improving leaderboard performance. However, we believe it is fair to use such statistics for the purpose of the current post hoc analysis.

We manually inspected a sample of matched passages to ensure that the matching procedure was reliable. Moreover, the distribution of positions of relevant passages matches that of a related FIRA dataset [2] (see Table 2), where such information was collected by crowdsourcing.

We also obtained passage-document matching data from the FIRA dataset [2], where fine-grained relevance information was crowdsourced for a small set of queries from TREC DL 2019. In both cases, we plot the distribution of *ending* positions of relevant passages (see Fig. 2a and Fig. 2b). When a document contains multiple relevant passages, we plot *only the first* one. Compared to the overall distribution of relevant document lengths, which has a long tail (see Fig. 1), the *first* relevant passage occurs typically in the first chunk. We hypothesize that such a skew makes it easy for the *FirstP* models to accurately rank documents. In the remaining of this sub-section we try to answer the following questions:

- What is the source of bias?
- Is there room for improvement by considering additional relevant passages beyond the first BERT chunks?

Position Bias of Relevant Passages. In the case of Robust04, unfortunately, little is known about the interface and annotation procedure used by relevance assessors. It is quite possible that they observed either complete documents or their starting parts: Judging complete long documents likely required scrolling. Thus, we conjecture that their perception of the relevance was likely biased towards the beginning of a document. In the case of the MS MARCO *dev* set, annotators judged each document passage independently. Then, passage-level relevance labels were transferred to the original documents.

Because annotators did not observe complete documents their attention was not biased towards the document start. Likewise annotators of the FIRA dataset (which uses TREC DL 2019 queries) judged randomly selected document snippets, which should have prevented attention bias. Furthermore, Hofstätter et al. [2] carried out additional experiments to confirm that this was the case. We conclude that the MS MARCO document collection has a content bias, but the exact nature of this bias is not clear. In summary, we want to emphasize that both Robust04 and MS MARCO, which are possibly the most commonly used retrieval datasets, are not particularly useful for benchmarking long-document models.

Considering Additional Relevant Passages. Although plots (see Fig. 1) provide some explanation, it is not a complete one. For a more accurate estimation of potential improvements from using longer document prefixes, we need a more detailed matching statistics as well as additional assumptions on model's capabilities to recognize relevant passages. Let us assume that when a relevant passage fully fits into the maximum supported document prefix (477), the average accuracy score, e.g. MRR, is maxed out at $C_{1\text{stp}}$. We also assume that the score is zero when the passage *starts* beyond the token number 477.

[1] The longest common subsequence relies on a sliding-window approach where the length of the window is 20% longer than the length of the passage we are trying to match.

Table 2: Distribution of start/end positions of relevant passages inside documents (chunks size is 477 BERT tokens)

input chunk #	development set (estimated)		FIRA (crowd-sourced)	
	start	end	start	end
1	85.9%	71.0%	83.8%	76.4%
2	9.1%	14.9%	9.9%	15.3%
3	2.6%	6.1%	2.3%	3.9%
4	1.2%	3.0%	2.2%	2.2%
5	0.6%	1.4%	0.7%	0.9%
6	0.6%	1.2%	0.4%	0.5%
6+	0.1%	2.5%	0.7%	0.7%

Based on our assumptions, according to Table 2, a *FirstP* model has a chance to "score" fully in about 71% of the cases on the MS MARCO v1 development set and in 76% on TREC 2019 DL queries.

At the same time, in about 10% of the cases the first relevant passage starts in the second chunk, which also means they end in the second or third one (MS MARCO passages are shorter than 477 tokens). According to our assumptions, the *FirstP* model should get a zero score for such documents. In contrast, our full-document models (which uses three chunks) could potentially achieve the score $1.1 \cdot C_{1\text{stp}}$, thus, outperforming *FirstP* by 10%. Yet, the actual improvements are only about 5%.

Furthermore, as many as 4-5% of *first* relevant passages end in chunks 4-6. However, as we can see from Table 1 increasing the prefix length beyond three chunks is not helpful on Robust04 (we observe the same on MS MARCO with Neural Model 1 [1]).

2 FULL EXPERIMENTAL REULTS

REFERENCES

[1] Leonid Boytsov and Zico Kolter. 2021. Exploring Classic and Neural Lexical Translation Models for Information Retrieval: Interpretability, Effectiveness, and Efficiency Benefits. In *ECIR (1) (Lecture Notes in Computer Science, Vol. 12656)*. Springer, 63–78.
[2] Sebastian Hofstätter, Markus Zlabinger, Mete Sertkan, Michael Schröder, and Allan Hanbury. 2020. Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering. In *CIKM*. ACM, 3031–3038.

Table 3: Model ranking performance averaged over seeds.

Model	MS MARCO	TREC DL				Robust04	
	dev	2019	2020	2021	2019-2021	title	description
	MRR	**NDCG@10**				**NDCG@20**	
AvgP	0.389^{abc}	0.659^{a}	0.596^{bc}	0.664^{c}	0.642^{c}	0.478^{bc}	0.531^{bc}
FirstP (BERT)	0.394^{bc}	0.631^{c}	0.598^{bc}	0.660^{c}	0.632^{bc}	0.475^{bc}	0.527^{bc}
FirstP (Longformer)	0.404^{abc}	0.657^{a}	0.616^{c}	0.654^{c}	0.643^{c}	0.483^{bc}	0.540^{c}
FirstP (ELECTRA)	$0.417^{a\ c}$	0.652^{c}	0.642^{a}	0.686^{a}	$0.662^{a\ c}$	$0.492^{a\ c}$	$0.552^{a\ c}$
MaxP	0.392^{bc}	0.648^{c}	0.615^{c}	0.665^{c}	$0.644^{a\ c}$	0.488^{abc}	0.544^{abc}
SumP	0.390^{bc}	0.642^{c}	0.607^{c}	0.662^{c}	0.639^{bc}	0.486^{bc}	0.538^{bc}
CEDR-DRMM	0.385^{abc}	0.639^{c}	0.592^{bc}	0.651^{bc}	0.629^{bc}	0.466^{bc}	0.533^{bc}
CEDR-KNRM	0.379^{abc}	0.637^{c}	0.599^{bc}	0.651^{bc}	0.630^{bc}	0.483^{bc}	0.535^{bc}
CEDR-PACRR	0.395^{bc}	0.640^{c}	$0.615^{a\ c}$	0.667^{c}	$0.643^{a\ c}$	$0.496^{a\ c}$	$0.549^{a\ c}$
Neural Model1	0.398^{bc}	0.660^{a}	$0.620^{a\ c}$	0.666^{c}	$0.650^{a\ c}$	0.484^{bc}	0.537^{bc}
PARADE Attn	$0.416^{a\ c}$	0.647^{c}	0.626^{a}	0.677^{c}	$0.652^{a\ c}$	$0.503^{a\ c}$	$0.556^{a\ c}$
PARADE Attn (ELECTRA)	$\mathbf{0.431}^{ab}$	0.675^{ab}	$\mathbf{0.653}^{a}$	$\mathbf{0.705}^{ab}$	$\mathbf{0.680}^{ab}$	$\mathbf{0.523}^{ab}$	$\mathbf{0.581}^{ab}$
PARADE Avg	0.392^{bc}	0.656^{a}	0.617^{c}	0.660^{bc}	$0.646^{a\ c}$	0.483^{bc}	0.534^{bc}
PARADE Max	0.405^{abc}	0.652^{c}	$0.626^{a\ c}$	$0.680^{a\ c}$	$0.655^{a\ c}$	0.489^{abc}	$0.548^{a\ c}$
PARADE Transf-RAND-L2	$0.419^{a\ c}$	0.657^{a}	$0.620^{a\ c}$	$0.681^{a\ c}$	$0.655^{a\ c}$	0.488^{abc}	$0.548^{a\ c}$
PARADE Transf-PRETR-L6	0.402^{abc}	0.646^{c}	0.608^{c}	0.667^{c}	0.643^{c}	0.494^{abc}	$0.554^{a\ c}$
PARADE Transf-PRETR-LATEIR-L6	0.398^{bc}	0.638^{c}	0.587^{bc}	0.649^{bc}	0.626^{bc}	0.450^{abc}	0.501^{abc}
LongP (Longformer)	$0.412^{a\ cd}$	$\mathbf{0.676}^{ab\ d}$	$0.628^{a\ c}$	$0.693^{a\ d}$	$0.668^{ab\ d}$	$0.500^{a\ cd}$	$0.568^{a\ d}$
LongP (Big-Bird)	0.397^{bc}	0.655^{c}	0.618^{c}	0.675^{c}	$0.651^{a\ c}$	0.452^{abc}	0.477^{abc}

Superscripts **a**, **b**, and **c** denote a statistical significant difference (at level 0.05) with respect to the following baselines: *FirstP (BERT)*, *PARADE Attn*, and *PARADE Attn (ELECTRA)*. The superscript **d** denotes a difference between LongP and FirstP variants of Longformer.