Investigating the Successes and Failures of BERT for Passage Re-Ranking

Harshith Padigela, Hamed Zamani, and W. Bruce Croft
College of Information and Computer Sciences
University of Massachusetts Amherst
Amherst, MA 01003
{hpadigela,zamani,croft}@cs.umass.edu

ABSTRACT

The bidirectional encoder representations from transformers (BERT) model has recently advanced the state-of-the-art in passage re-ranking. In this paper, we analyze the results produced by a fine-tuned BERT model to better understand the reasons behind such substantial improvements. To this aim, we focus on the MS MARCO passage re-ranking dataset and provide potential reasons for the successes and failures of BERT for retrieval. In more detail, we empirically study a set of hypotheses and provide additional analysis to explain the successful performance of BERT.

1 INTRODUCTION

Recent developments in deep learning and the availability of large-scale datasets have led to significant improvements in various computer vision and natural language processing tasks. In information retrieval (IR), the lack of publicly available large-scale datasets for many tasks, such as ad-hoc retrieval, has restricted observing substantial improvements over traditional methods [6, 13]. A number of approaches, such as weak supervision [2, 12], have been recently proposed to enable deep neural models to learn from limited training data. More recently, Microsoft has released MS MARCO v2 [8], a large dataset for the passage re-ranking task, to foster the neural information retrieval research.

In this paper, we first show that a simple neural model that uses bidirectional encoder representations from Transformers (BERT) [3] for question and passage representations performs surprisingly well compared to state-of-the-art retrieval models, including traditional term-matching models, conventional feature-based learning to rank models, and recent neural ranking models. This has been also discovered by other researchers, such as [9] in parallel with this study. Looking at the leaderboard of the MS MARCO passage re-ranking task shows the effectiveness of the BERT representations for retrieval.[1]

We believe that understanding the performance of effective neural IR models, e.g., BERT, is important. It could potentially provide guidelines for the IR researchers for further development of neural IR models. Given this motivation, this paper mainly analyzes the results obtained by BERT for passage re-ranking and studies the reasons behind its success. To do so, we compare the results obtained by both BM25 and BERT, and highlight their differences. We choose BM25 as our basis for comparison, due to its effectiveness and more importantly its simplicity and explainable behavior, which makes the analysis easier.

In more detail, this paper studies the following hypotheses:

- H1: BM25 is more biased towards higher query term frequency compared to BERT.
- H2: Bias towards higher query term frequency hurts the BM25 performance.
- H3: BERT retrieves documents with more novel words.
- H4: BERT's improvement over BM25 is higher for longer queries.

In addition we also identify the query types for which BERT does and does not perform well. Our experiments provide interesting insights into the performance of this model.

2 BERT

Representations learned using language modelling [7, 10] have shown to be useful in many downstream natural language tasks [1]. There exist two primary approaches for using these pre-trained representations: (1) feature-based models and (2) fine-tuning [3]. In the feature-based approach, task-specific architectures are designed on top of the pre-trained feature representations. While in the fine-tuning approach, minimal task specific parameters are added, which will be fine-tuned in addition to the pre-trained representations for the downstream task. BERT [3], which falls into the latter category, is a multi-layer bidirectional transformer encoder utilizing the transformer units described in [11]. The BERT model uses bidirectional self-attention to capture interaction between the input tokens and is pre-trained on the masked language modelling task [3].

Pre-trained BERT models, fine-tuned using a single additional layer have been shown to achieve state-of-the-art results in a wide range of natural language tasks, including machine reading comprehension (MRC) and natural language inference (NLI) [3]. In this paper, we also use the same setting, by adding a single layer on top of the BERT's large pre-trained model, and fine-tuning it using a pointwise setting with a maximum likelihood objective. This is also similar to the setting used in [9].

[1] The leaderboard is available at http://www.msmarco.org/leaders.aspx.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 EMPIRICAL ANALYSIS

3.1 Data

We consider the MS MARCO dataset for passage re-ranking [8] in our analysis. The MS MARCO dataset is generated using queries sampled from the Bing's query logs and corresponding relevant passages marked by human assessors. It is notable that the relevance judgments provided by the MS MARCO dataset are different from the traditional TREC-style relevance judgments. They utilized the information that the human assessors provided for the machine reading comprehension data. This means that a marked passage is a true positive/relevant, however an unmarked passage may not be a true negative. For every query, a set of top 1000 candidate passages are extracted using BM25 for re-ranking. Since the original relevant documents were picked from a set of 10 candidate documents chosen by the Bing's ranking stack, the relevant passages might not be present in the 1000 passages chosen by BM25.

The training set consists of approximately 400 million tuples of query, relevant passage, and non-relevant passage. The development set contains 6,980 queries with their corresponding set of 1,000 candidate passages. On average, each query has one relevant passage. Around 1242 queries have no marked relevant passages. We primarily focus our analysis only on the 5738 queries in the development set that have at least one relevant passage.

3.2 Experimental Setup

We use BERT and BM25 for our analysis and comparison. We used the BERT large model trained on MS MARCO. The training setup is similar to the one described in [9]. For BM25 relevance matching we indexed all the passages using Elasticsearch [4] with the default analyzer and default parameters of $b = 0.75$ and $k1 = 1.2$. Since most queries have only 1 relevant document, we use mean reciprocal rank of the top 10 retrieved passages (MRR@10) as our main retrieval metric, which is also suggested by MS MARCO [8].[2]

3.3 Results and Discussion

The performance of various models on the entire development and evaluation (test) sets of MS MARCO are shown in Table 1. We can see that the BERT model which was originally trained on the masked language modelling (MLM) task [3] and further fine-tuned using a pointwise training on the MS MARCO data, outperforms the existing traditional retrieval models and recent neural ranking models by a large margin. In order to understand these improvements, we look into the BERT's and the BM25's performances on the development set.[3]

In the following, we study a set of hypotheses and provide empirical evidence to either validate or invalidate them.

Hypothesis I: BM25 is more biased towards higher query term frequency compared to BERT. We hypothesize that in many queries the top results from BM25 were just the passages that contain multiple repetitions of query words without actually conveying any useful information, which is not the case for BERT.

[2]Due to the incomplete judgments, recall-oriented metrics such as mean average precision (MAP), are not suitable for this dataset.
[3]Note that the evaluation set is not publicly accessible.

Table 1: MRR@10 percentage from the MS MARCO leaderboard.

Model	Eval	Dev
BM25	16.49	16.70
BM25 (ours)	-	17.67
Feature-based LeToR: with RankSVM	19.05	19.47
Neural Kernel Match IR (KNRM)	19.82	21.84
Neural Kernel Match IR (Conv-KNRM)	27.12	29.02
IRNet (Deep CNN/IR Hybrid Network)	28.06	27.80
BERT + Small Training	**35.87**	**36.53**

Table 2: Average MRR and # of queries (in parenthesis) for ranges of FQT.

FQT	[0, 0.1)	[0.1, 0.15)	[0.15, 0.2)	[0.2, 0.25)	[0.25, 1]
BM25	0.29 (349)	0.23 (1163)	0.22 (1565)	0.20 (1316)	0.19 (1345)
BERT	0.48 (1240)	0.47 (2061)	0.42 (1441)	0.38 (652)	0.40 (344)

To validate this hypothesis, we calculate the *fraction of query tokens (FQT)* as follows: For each query, we take the top k results, remove stopwords and punctuations, and calculate the fraction of query tokens in the remaining tokens. If $d_1, d_2, \cdots, d_k$ are the set of results for a query q without stopwords and punctuation, then,

$$FQT(q) = \frac{1}{k} \sum_{i=1}^{k} \frac{N(d_i, q)}{|d_i|} \quad (1)$$

where $N(d_i, q)$ denotes the number of occurrences of query tokens q in the document d_i. We limit k to a maximum of 10.

We find that the FQT average across queries is 0.2 for BM25 and 0.147 for BERT. In 95.96% of the queries BM25 has a higher FQT value than BERT. These results validate our first hypothesis, saying that BM25 has a higher bias towards query term frequency in document matches, compared to BERT. An example can be seen in Table 5 Query 1

Hypothesis II: Bias towards higher query term frequency hurts the BM25 performance. We hypothesize that the bias towards query term frequency affects the BM25 performance significantly, compared to BERT.

To investigate this, we see how MRR changes across different ranges of FQT. The FQT range of [0,1] is split into 5 buckets and the average MRR value and the number of queries in each bucket is shown in Table 2. As FQT value increases, we can see that the MRR value decreases in both BERT and BM25. Because of BM25's bias towards high FQT (validated by Hypothesis I and also evident by the number of queries), we can see the decrease in MRR (as we go from the lowest to highest FQT buckets) is more prominent for BM25 (34.5%) than BERT (16.7%). The signed t-test for measuring the difference between two pairs of data, applied on difference between FQT values of BM25 and BERT yields a p-value of 0.0, indicating statistically significant difference between the FQT values.

Hypothesis III: BERT retrieves documents with more novel words. Since the recent neural models trained on the language modeling task have been shown to capture semantic similarities,

Table 3: Average MRR with respect to query length (L).

L	2	3	4	5	6	7	8	9	10
BM25	0.27	0.23	0.22	0.22	0.23	0.19	0.21	0.17	0.18
BERT	0.56	0.46	0.48	0.45	0.46	0.42	0.40	0.38	0.34

Table 4: Average MUR with respect to the different cut-off values (i).

i	1	2	3	4	5	6	7	8	9	10
Avg. MUR	0.17	0.45	0.77	1.1	1.44	1.78	2.13	2.46	2.8	3.12

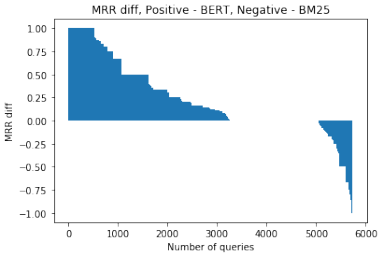

Figure 1: $MRR_{BERT} - MRR_{BM25}$ on MSMarco Dev set.

we hypothesize that BERT can retrieve results with more novel words, compared to BM25. To validate this, we calculate the *fraction of novel terms (FNT)* as follows. Let $d_1, d_2, \cdots, d_k$ be the results for a query q, which are stripped of stopwords and punctuation. Then

$$FNT(q) = \frac{1}{k} \sum_{i=1}^{k} \frac{N'(d_i, q)}{U(d_i)} \tag{2}$$

where $U(d_i)$ gives the number of unique terms in document d_i and $N'(d_i, q)$ gives the number of unique terms in document d_i which are not present in the query q. We limit k to a maximum of 10. We find that the FNT average across queries is 0.88 for BM25 and 0.9 for BERT. In 85.85% of queries BERT has a higher FNT value than BM25. The signed t-test on the difference between FNT values of BERT and BM25 yields a p-value of 0.0, indicating statistically significant difference between the FNT values. This validates our hypothesis that BERT retrieves documents with more novel words than BM25.

Hypothesis IV: BERT's improvement over BM25 is higher for longer queries. Since the BERT model is designed to learn context-aware word representations, we hypothesize that its improvements for longer queries, which generally provide richer context, are more significant. To validate this hypothesis, we calculate the average MRR per query length for both BERT and BM25, shown in Table 3. We can see that BERT performs significantly better than BM25 across all query lengths. But as the query length increases from 2 to 10, the performance of both BM25 and BERT generally decreases, and this decrease is more prominent for BERT (39%) compared to BM25 (33%), indicating its higher sensitivity to query length than BM25. The MRR difference between BERT and BM25 also decreases from 0.29 to 0.16 as query length increases from 2 to 10, which indicates that our fourth hypothesis is incorrect and that BERT's improvement is lower for longer queries. Interestingly, BERT performs surprisingly well for very short queries compared to longer ones. The reason might be that BERT is not successful at capturing the query context properly for long queries. This can be also observed from the examples, such as Queries 3, 4 in Table 5.

3.4 Result Analysis

We conduct various analyses to understand the similarities and differences between BM25 and BERT. We discuss them below.

Per Query Analysis. We analyze the per query performance of BERT compared to BM25. Figure 1 plots ΔMRR per query (i.e., $MRR_{BERT} - MRR_{BM25}$), which are sorted in descending order. As depicted in the figure, in 3257 questions (57% out of 5738) BERT has a better performance compared to BM25 and in 690 questions (12%) BM25 performs better than BERT. For 525 (9%) queries ΔMRR is equal to 1, meaning that a relevant answer is retrieved by BERT as the first ranked passage, however, no relevant answer is retrieved by BM25 in the top 10 result list. For 1791 queries, both BERT and BM25 perform similarly.

This experiments show that BERT not only performs better than BM25 (on average), but it also performs more accurately for substantially more queries.

Similarity between the BERT's and the BM25's result lists. To measure the similarity between the results of BERT and BM25, we calculate the following metric, MUR - matches upto result. MUR(i, q) for a query q measures the number of matches in the top i results of BERT and BM25. We can see the average MUR for each $i \in [1, 10]$ in Table 4 which indicates the low extent of similarity between BERT and BM25. The number of matches increases linearly with i with slope of about 0.33 and intercept around -0.21, indicating a consistent linear relationship between BERT and BM25.

Comparison by answer type. In order to understand the performance of these models across different types of questions, we classify questions based on the lexical answer type. We use the rule-based answer type classifier[4] inspired by [5] to extract answer types. We classify questions based on 6 answer types, namely abbreviation, location, description, human, numerical and entity. The average MRR across these 6 types for 4105 queries (having a valid answer type) is shown in Table 6. We can see that while BERT has highest MRR on *abbreviation* type questions, BM25 has its lowest MRR on them. Note that BERT seems to have its lowest performance on *numerical* and *entity* type questions.

Comparison using query starting ngrams: Here we look at the most frequent bigrams with which the queries start. The idea is that looking at the starting ngrams can help us understand the type of queries. We extract the most frequent 15 bigrams and compute the average MRR using BERT for each of them. The result is shown in Figure 2. We can see that the bigrams corresponding to *numeric* type questions, such as "how much" and "how long", as well as *location* type questions like "what county / where is" and *entity* type questions, such as "what type" have a low MRR. This is consistent with our observations in the previous experiment (see Table 6).

Semantic similarity: Being trained on a language modeling task, we expect BERT to capture various semantic relationships. While in some cases these help in arriving at the right answer sometimes

[4]https://github.com/superscriptjs/qtypes

Table 5: Sample queries for comparison. (W) - incorrect and (C) - correct result.

ID	Query	BM25/BERT/Relevant passage	Comparison
1	what is the nationality	*BM25*: Users found this page by searching for 1 is african american a nationality 2 african american nationality 3 is black a nationality nationality african .... *BERT*: Nationality is the legal relationship between a person and a nation state	BM25(W) vs BERT(C)
2	confident man definition	*BM25*: definition of suave is someone smooth confident .... usually describing a man *BERT*: definition of a confidence man is someone who gets a victim to trust them before taking their money or property, a con man	BM25(W) vs BERT(C)
3	where can a plasma membrane be found in a cell	*BERT*: The Plasma membrane is found in both the animal cell and plant cell *Rel*: The plasma membrane is the border between the interior and exterior of a cell	BERT(W) vs Relevant
4	telephone number for amazon fire stick customer service	*BERT*: Customer Service 1 866 216 1072. .... 1 Thank you for calling Amazon.com customer service for quality assurance and training .... *Rel*: .... for more information contact amazon fire stick support number 1 8447451521	BERT(W) vs Relevant
5	another name for reaper	*BERT*: Reaper orginally known as Gabriel Reyes is a mercenary.... is antagonist .... in videogame Overwatch. He is voiced by Keith Ferguson who also played Lord Hater *Rel*: .... similar words for the term reaper. harvester reaper	BERT(W) vs Relevant

Table 6: Average MRR values for answer types. Sorted by ΔMRR.

Type	ABBR	LOC	DESC	HUM	NUM	ENTY
# queries	9	493	1887	455	933	328
BM25	0.17	0.25	0.19	0.23	0.19	0.21
BERT	0.59	0.50	0.43	0.46	0.40	0.41

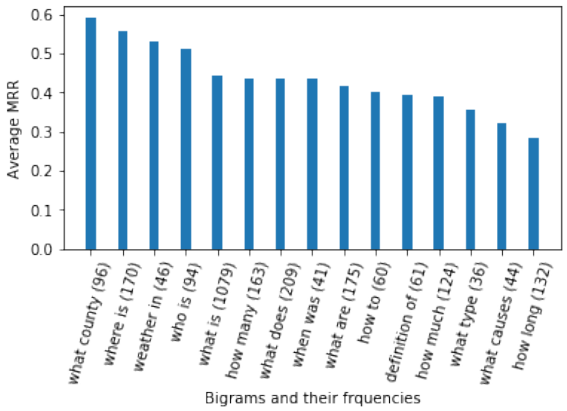

Figure 2: Average MRR of Frequent Bigrams.

they can also lead to incorrect answers. We will discuss two such examples below. In question 2 of Table 5, BERT captures similarity between the word "confident" in query and "confidence" in the passage, which helps it arrive at the right answer. This can be seen by visualizing the attention values between query and document words as shown in Figure 3. Similarly in Example 5 of Table 5, the question asks for another *name* for word "reaper", which in this context means synonyms for the word "reaper". However, BERT relates *name* to a *character name* reaper (see attention map 4). This leads to an incorrect answer.

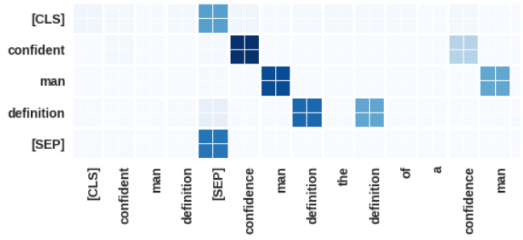

Figure 3: Attention map of head 14 from BERT layer 16.

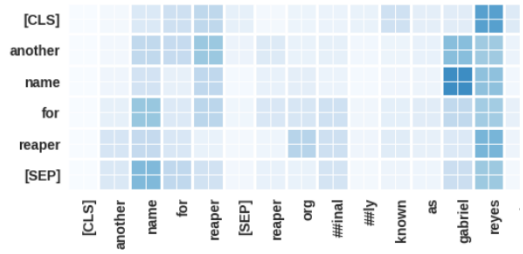

Figure 4: Attention map of head 4 from BERT layer 23.

4 CONCLUSIONS AND FUTURE WORK

BERT performs surprisingly well for a passage re-ranking task. In this paper, we provide empirical analysis to understand the performance of BERT and how its results are different from a typical retrieval model, e.g., BM25. We showed that BM25 is more biased towards high query term frequency and this bias hurts its performance. We demonstrated that, as expected, BERT retrieves passages with more novel words. Surprisingly, we found out that BERT is failing at capturing the query context for long queries. Our analysis also suggested that BERT is relatively successful in answering *abbreviation* answer type questions and relatively poor at *numerical* and *entity* type questions.

Although BERT substantially outperforms state-of-the-art models for passage retrieval, it is still far away from a perfect retrieval performance. We believe that future work investigating the relevance preferences captured by BERT across various query types

and a better encoding of query context for longer queries could help in developing even better models.

5 ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF IIS-1715095. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR* 12, Aug (2011), 2493–2537.

[2] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. *SIGIR* (2017), 65–74.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805

[4] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine.* " O'Reilly Media, Inc.".

[5] Xin Li and Dan Roth. 2002. Learning question classifiers. *COLING* (2002), 1–7.

[6] Jimmy Lin. 2019. The Neural Hype and Comparisons Against Weak Baselines. *SIGIR Forum* 52, 2 (Jan. 2019), 40–51.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *NIPS* (2013), 3111–3119.

[8] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268

[9] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *CoRR* abs/1901.04085 (2019). arXiv:1901.04085

[10] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. *EMNLP* (2014), 1532–1543.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*. 5998–6008.

[12] Hamed Zamani and W Bruce Croft. 2018. On the theory of weak supervision for information retrieval. *ICTIR* (2018), 147–154.

[13] Hamed Zamani, Mostafa Dehghani, Fernando Diaz, Hang Li, and Nick Craswell. 2018. SIGIR 2018 workshop on learning from limited or noisy data for information retrieval. *The 41st International ACM SIGIR Conference* (2018), 1439–1440.