Diagnosing BERT with Retrieval Heuristics

Arthur Câmara and Claudia Hauff

Delft University of Technology
Delft, the Netherlands
{a.barbosacamara,c.hauff}@tudelft.nl

Abstract. Word embeddings, made widely popular in 2013 with the release of word2vec, have become a mainstay of NLP engineering pipelines. Recently, with the release of BERT, word embeddings have moved from the term-based embedding space to the contextual embedding space—each term is no longer represented by a single low-dimensional vector but instead each term and *its context* determine the vector weights. BERT's setup and architecture have been shown to be general enough to be applicable to many natural language tasks. Importantly for Information Retrieval (IR), in contrast to prior deep learning solutions to IR problems which required significant tuning of neural net architectures and training regimes, "vanilla BERT" has been shown to outperform existing retrieval algorithms by a wide margin, including on tasks and corpora that have long resisted retrieval effectiveness gains over traditional IR baselines (such as Robust04). In this paper, we employ the recently proposed axiomatic dataset analysis technique—that is, we create diagnostic datasets that each fulfil a retrieval heuristic (both term matching and semantic-based)—to explore what BERT is able to learn. In contrast to our expectations, we find BERT, when applied to a recently released large-scale web corpus with ad-hoc topics, to *not* adhere to any of the explored axioms. At the same time, BERT outperforms the traditional query likelihood retrieval model by 40%. This means that the axiomatic approach to IR (and its extension of diagnostic datasets created for retrieval heuristics) may in its current form not be applicable to large-scale corpora. Additional—different—axioms are needed.

1 Introduction

Over the course of the past few years, IR has seen the introduction of a large number of successful deep learning approaches for solving all kinds of tasks previously tackled with hand-crafted features (within the learning to rank framework) or traditional retrieval models such BM25.

In 2017, with the introduction of the transformer architecture [33], a second wave of neural architectures for NLP has emerged. Approaches (and respective models) like BERT [7], XLNet [41] and GPT-2 [24] have shown that it is indeed possible for one general architecture to achieve state-of-the-art performance across very different NLP tasks (some of which are also related to IR tasks, such as question answering, reading comprehension, etc.).

Ad-hoc retrieval, the task of ranking a set of documents given a single query, has long resisted the success of neural approaches, especially when employed across standard IR test collections such as Robust04[1], which come with hundreds of topics (and thus relatively little training data). Often, the proposed neural approaches require a very careful design of their architecture. Also, the training regime and the input data transformations have to be *just right* [17] in order to beat or come close to well-tuned traditional IR baselines such as RM3 [16,2][2]. With the introduction of BERT in late 2018, this has finally changed. Recently, a range of BERT-inspired approaches have been shown to clearly surpass all strong IR baselines on Robust04 [6,40] and other IR corpora.

It is still an open question though what exactly makes BERT and similar approaches perform so well on IR tasks. While recent works try to understand what BERT learns most often by analysing attention values, e.g., [32,21,12,4], analysing BERT under the IR light requires a different set of tools. While most NLP tasks optimise for precision, recall or other objective metrics, the goal of ad-hoc retrieval is to optimise for *relevance*, a complex multidimensional and somewhat subjective concept [3].

In this paper, we set out to explore BERT under the IR lens, employing the concept of *diagnostic datasets*, an IR model analysis approach (inspired by similar NLP and computer vision approaches) proposed last year by Rennings et al. [25]. The idea behind these datasets is simple: each dataset is designed to fulfil one *retrieval axiom* [9], i.e., a heuristic that a good retrieval function should fulfil[3]. Each dataset contains query-documents instances (most often, a query and two documents) that the investigated model should rank in the correct order as determined by the heuristic. The extent to which a model correctly ranks those instances is allowing us to gain insights into what type of information the retrieval model pays attention to (or not) when ranking documents. While traditional retrieval models such as BM25 [26] can be analysed analytically, neural nets with their millions or even billions of learnt weights can only be analysed in such an empirical manner.

More concretely, we attempt to analyse a version of BERT, DistilBERT (that was shown to attain 97% of "vanilla" BERT performance [28]), fine-tuned on the TREC 2019 Deep Learning track dataset[4]. We extend previous work [25] by incorporating additional axioms (moving from term matching to semantic axioms). We find that DistilBERT to outperform the traditional query likelihood (QL) model by 40%. In contrast to our expectations however, we find that BERT does not adhere to any of the axioms we incorporate in our work. This implies that the currently existing axioms are *not sufficient* and *not applicable* to capture

[1] Robust04 is a test collection employed at the TREC 2004 robust retrieval task [35], consisting of 528K newswire documents, 250 topics and 311K relevance judgements.

[2] We want to emphasise here that this observation is specific to IR corpora with few training topics; for the very few corpora with hundreds of thousands of released topics (such as MSMarco) this observation does not hold.

[3] As a concrete example, consider the TFC1 [9] heuristic: *The more occurrences of a query term a document has, the higher its retrieval score.*

[4] https://microsoft.github.io/TREC-2019-Deep-Learning/

the heuristics that a strong supervised model learns (at least for the corpus and model we explore); it is not yet clear to what extent those results generalise beyond our model and corpus combination but it opens up a number of questions about the axiomatic approach to IR.

2 Related Work

Axiomatic Information Retrieval The use of axioms (or "retrieval heuristics") as a means to improve and understand information retrieval techniques is well established. It is an analytic technique to explore retrieval models and how best to improve them. In their seminal work, Fang et al. [9,10] introduced a number of term-matching based *heuristics* that models should follow in order to be successful in retrieval tasks. Subsequently, Fang et al. [11] proposed a set of axioms based on semantic matching and thus allowing non-exact matches to be accounted for in axiomatic retrieval. We apply these axioms in our work—albeit in a slightly adapted manner. Other applications for axioms in IR include document re-ranking based on a Learning to Rank scenario [13] and query expansion [8] by exploring similar axioms. It should be noted, that—while sensible—it cannot be assumed that these axioms are a good fit for all kinds of corpora; they represent a general notion of how a good retrieval function should behave. Recently, Rennings et al. [25] introduced *diagnostic datasets* extracted from actual corpora that each fulfil one axiom. In contrast to the axiomatic approach, which requires an analytical evaluation of the retrieval functions under investigation, a diagnostic corpus enables us to analyse models' axiomatic performance that are too large to be analysed analytically (such as neural models with millions or even billions of parameters[5]). Our work continues in that direction with a larger number of axioms (9 vs. 4) and the analysis of the current neural state-of-the-art (i.e., BERT).

Neural IR Models Neural IR models, i.e., deep learning based approaches that tackle IR problems, have seen a massive rise in popularity in the last few years, with considerable success [20]. Models like DRMM [19], ARCII [14] and aNMM [39] have been shown to be suitable for a range of IR tasks when sufficient training data is available; it remains at best unclear at smaller data scale whether the reported successes are not just an artefact of weak baselines [17].

Recently, a new wave of approaches, based on the transformer architecture [33] has shown that, finally, neural models can significantly outperform traditional and well-tuned retrieval methods such as RM3 [2]. Yang et al. [40] have shown that BERT, fine-tuned on the available TREC microblog datasets, and combined with a traditional retrieval approach such as query likelihood significantly outperforms well-tuned baselines, even on Robust04 which has shown to be a notoriously difficult dataset for neural models to do well on. With similar success, Dai and Callan [6] have recently employed another BERT variant

[5] As a concrete example, our BERT model contains 66 million parameters.

on Robust04 and ClueWeb09. Lastly we point out, that works are now also beginning to appear, e.g., [18], that use the contextual word embeddings produced by BERT in combination with another strong neural model, again with strong improvements over the existing baselines.

Analysing Neural IR Models As we aim to analyse BERT, we also consider how others have tackled this problem. Analysing neural models—whether for IR, NLP or another research domain—is not a trivial task. By now a great number of works have tried to light up the black box of the neural learning models [1], with varying degrees of success. Within IR, Pang et al. [22] have aimed to paint a complete and high-level picture of the neural IR area, comparing the behaviour of different approaches, and showing that interaction and representation-based models focus on different characteristics of queries and documents. While insightful, such work does not enable us to gain deep insights into a single model. Closer to our work, Rosset et al. [27] employ axioms to generate artifical documents for the training of neural models and the regularization of the loss function. In contrast, we employ axioms to *analyze* retrieval models.

Another direction of research has been the development of interpretation tools such as DeepSHAP [12] and LIRME [34] that aim to generate *local* explanations for neural IR models. Recently, in particular BERT (due to its successes across a wide range of tasks and domains) has become the focus of analysis—not within IR though. While approaches like [4] explore the attention values generated by the model's attention layers, Tenney et al. [32] argue that BERT is re-discovering classical NLP pipeline approaches in its layers, *"in an interpretable and localizable way"*, essentially repeating traditional NLP steps in a similar order as an expert would do, with steps like POS tagging, parsing, NER and coreference resolution happening within its layers in the expected order. Finally, Niven et al. [21] raise some critical points about BERT, arguing that it only *"exploits spurious statistical cues in the dataset"*; they showcase this by creating adversarial datasets that can significantly harm its performance.

3 Diagnostic Datasets

The usage of diagnostic datasets as a means to analyse neural models is common in NLP, e.g. [37,15,36] as there are a large number of fine-grained linguistic tasks (anaphora resolution, entailment, negation, etc.) that datasets can be created for with relative ease. In contrast, in IR the central notion is relevance and although we know that it can be decomposed into various types (topical, situational, etc.) of relevance [29], we have no easy way of creating datasets for each of these—it remains a time-intensive and expensive task. This also explains why corpora such as Robust04 remain useful and in use for such a long time. Instead, like Rennings et al. [25] we turn to the axiomatic approach to IR and create diagnostic datasets—one for each of our chosen retrieval heuristics. It has been shown that, generally, retrieval functions that fulfil these heuristics achieve a greater effectiveness than those that do not. In contrast to [25] which restricted itself to four term matching axioms, we explore a wider range of axioms, covering term frequency, document length, lower-bounding term frequency, semantic term matching and term proximity constraints. In total, we explore 9 axioms, all of which are listed in Table 1 with a short informal description of their main assumption of what a sensible retrieval function should fulfil. We note that this covers most of the term-matching and semantic-matching axioms that have been proposed. We have eliminated a small number from our work as we do not consider them relevant to BERT (e.g., those designed for pseudo-relevance feedback [5]).

As the axiomatic approach to IR has been designed to *analytically* analyse retrieval functions, in their original version they assume very specific artificial query and document setups. As a concrete example, let us consider axiom `STMC1` [11]. It is defined as follows: *given a single-term query* $Q = \{q\}$ *and two single-term documents* $D_1 = \{d_1\}$, $D_2 = \{d_2\}$ *where* $d_1 \neq d_2 \neq q$, *the retrieval score of* D_1 *should be higher than* D_2 *if the semantic similarity between* q *and* d_1 *is higher than that between* q *and* d_2. This description is sufficient to mathematically analyse classic retrieval functions, but not suitable for models with more than a handful of parameters. We thus turn to the creation of datasets that *exclusively* contain instances of query/documents (for `STMC1` an instance is a triple, consisting of one query and two documents) that satisfy a particular axiom. As single-term queries and documents offer no realistic test bed, we *extend* (moving beyond single-term queries and documents) and *relax* (moving beyond strict requirements such as equal document length) the axioms in order to extract instances from existing datasets that fulfil the requirements of the extended and relaxed axiom. Importantly, this process requires no relevance judgements—we can simply scan all possible triples in the corpus (consisting of queries and documents) and add those to our diagnostic dataset that fulfil our requirements. We then score each query/document pair with our BERT model[6] and determine whether the score order of the documents is in line with the axiom. If it is, we consider our model to have classified this instance correctly.

While Table 1 provides an informal overview of each heuristic, we now formally describe each one in more detail. Due to the space limitations, we focus on a mathematical notation which is rather brief. For completeness, we first state the original axiom and then outline how we extend and relax it in order to create a diagnostic dataset from it. For axioms `TFC1`, `TFC2`, `LNC2` and `M-TDC` we follow the process described in [25]. We make use of the following notation: Q is a query and consists of terms $q_1, q_2, ...$; D_i is a document of length $|D_i|$ containing terms $d_{i_1}, d_{i_2}, ...$; the count of term w in document D is $c(w, D)$; lastly, $S(Q, D)$ is the retrieval score the model assigns to D for a given Q. Apart from the proximity heuristic `TP`, the remaining heuristics are based on the bag-of-word assumption, i.e., the order of terms in the query and documents do not matter.

[6] Note, that scoring each document independently for each query is an architectural choice, there are neural architectures that take a query/doc/doc triplet as input and output a preference score.

TFC1—Original Assume $Q = \{q\}$ and $|D_1| = |D_2|$. If $c(q, D_1) > c(q, D_2)$, then $S(Q, D_1) > S(Q, D_2)$.

TFC1—Adapted In order to extract query/document/document triples from actual corpora, we need to consider multi-term queries and document pairs of approximately the same length. Let $Q = \{q_1, q_2, .., q_{|Q|}\}$ and $|D_1| - |D_2| \leq abs(\delta)$. $S(Q, D_1) > S(Q, D_2)$ holds, when D_1 has at least the same query term count as D_2 for all but one query term (and for this term D_1's count is higher), i.e., $c(q_i, D_1) \geq c(q_i, D_2)$ $\forall q_i \in Q$ and $\sum_{q_i \in Q} c(q_i, D_1) > \sum_{q_i \in Q} c(q_i, D_2)$.

TFC2—Original Assume $Q = \{q\}$ and $|D_1| = |D_2| = |D_3|$. If $c(q, D_1) > 0$, $c(q, D_2) - c(q, D_1) = 1$ and $c(q, D_3) - c(q, D_2) = 1$, then $S(Q, D_2) - S(Q, D_1) > S(Q, D_3) - S(Q, D_2)$.

TFC2—Adapted Analogous to `TFC1`, queries can contain multiple terms and documents only have to have approximately the same length. Let $Q = \{q_1, q_2, .., q_{|Q|}\}$ and $max_{D_i, D_j \in \{D_1, D_2, D_3\}}(|D_i| - |D_j| \leq abs(\delta))$. If every document contains at least one query term, and D_3 has more query terms than D_2 and D_2 has more query terms than D_1, and the difference of query terms count between D_2 and D_1 should be the same as between D_3 and D_2, for all query terms, i.e. $\sum_{q \in Q} c(q, D_3) > \sum_{q \in Q} c(q, D_2) > \sum_{q \in Q} c(q, D_1) > 0$ and $c(q, D_2) - c(q, D_1) = c(q, D_3) - c(q, D_2) \forall q \in Q$, then $S(Q, D_2) - S(Q, D_1) > S(Q, D_3) - S(Q, D_2)$.

M-TDC—Original Let $Q = \{q_1, q_2\}$, $|D_1| = |D_2|$, $c(q_1, D_1) = c(q_2, D_2)$ and $c(q_2, D_1) = c(q_1, D_2)$. If $idf(q_1) \geq idf(q_2)$ and $c(q_1, D_1) > c(q_1, D_2)$, then $S(Q, D_1) \geq S(Q, D_2)$.

M-TDC—Adapted Again, Let Q contain multiple terms and $|D_1| - |D_2| \leq abs(\delta)$. D_1, D_2 also differ in at least one query term count ($\exists q_i \in Q,$ such that $c(q_i, D_1) \neq c(q_i, D_2)$). If, for all query term pairs the conditions hold that $c(q_i, D_1) \neq c(q_j, D_j)$, $idf(q_i) \geq idf(q_j)$, $c(q_i, D_1) = c(q_j, D_2)$, $c(q_j, D_1) = c(q_i, D_2)$, $c(q_i, D_1) > c(q_i, D_2)$ and $c(q_i, Q) \geq c(q_j, Q)$, then $S(Q, D_1) \geq S(Q, D_2)$.

LNC1—Original Let Q be a query and D_1, D_2 be two documents. If for some $q' \notin Q$, $c(q', D_2) = c(q', D_1) + 1$ and for any $q \in Q$ $c(q, D_2) = c(q, D_1)$, then $S(Q, D_1) \geq S(Q, D_2)$.

LNC1—Adapted The axiom can be used with no adaptation.

LNC2—Original Let Q be a query. $\forall k > 1$, if D_1 and D_2 are two documents such that $|D_1| = k \cdot |D_2|$, and $\forall q \in Q, c(q, D_1) = k \cdot c(q, D_2)$, then $S(Q, D_1) \geq S(Q, D_2)$.

LNC2—Adapted The axiom can be used with no adaptation.

TP—Original Let $Q = \{q_1, q_2, ...q_{|Q|}\}$ be a query and D' a document generated by switching the position of query terms in D. Let $\sigma(Q, D)$ be a function that measures the distance of query terms $q_i \in Q$ inside a document D. If $\sigma(Q, D) > \sigma(Q, D')$, then $S(Q, D) < S(Q, D')$.

TP—Adapted Let $\mathit{\Gamma}(D, Q) = min_{(q_1,q_2 \in Q \cap D, q_1 \neq q_2)} Dis(q_1, q_2; D)$ be a function that computes the minimum distance between every pair of query terms in D. If $\mathit{\Gamma}(D_1, Q) < \mathit{\Gamma}(D_2, Q)$, then $S(Q, D_1) > S(Q, D_2)$.

For the following semantic axioms, let us define the function $\sigma(t_1, t_2)$ as the cosine distance between the embeddings of terms t_1 and t_2. We also define $\sigma'(T_1, T_2)$, where T can be either a document D or a query Q, as an extension to σ, defined by $\sigma'(T_1, T_2) = cos(\frac{\sum_{i \in T_1} t_i}{|T_1|}, \frac{\sum_{i \in T_2} t_i}{|T_2|})$, the cosine distance between the average term embeddings for each document.

STMC1—Original Let $Q = \{q\}$ be a one-term query, $D_1 = \{d_{1_1}\}$ and $D_2 = \{d_{2_1}\}$ be two single term documents, such that $d_{1_1} \neq d_{2_1}$, $q \neq d1_1$ and $q \neq d_{2_1}$. If $\sigma(q, d_{1_1}) > \sigma(q, d_{2_1})$, then $S(Q, D_1) > S(Q, D_2)$.

STMC1—Adapted We allow D_1 and D_2 to be arbitrarily long, covering the same number of query terms (i.e. $|D_1 \bigcap Q| = |D_2 \bigcap Q|$). Assume $\{D_i\} - \{Q\}$ be the document D_i without query terms, If $\sigma'(\{D_1\} - \{Q\}, Q) > \sigma'(\{D_2\} - \{Q\}, Q)$, then $S(Q, D_1) > S(Q, D_2)$.

STMC2—Original Let $Q = \{q\}$ be a one-term query and d a non-query term such that $\sigma(d, q) > 0$. If $D_1 = \{q\}$ and $|D_2| = k, (k \geq 1)$, composed entirely of d's (i.e.v$c(d, D_2) = k$), then $S(Q, D_1) \geq S(Q, D_2)$.

STMC2—Adapted We allow Q to be a multiple query term, D_1 to contain non-query terms and D_2 to contain query terms. If $\sum_{t_i, t_i \notin Q} c(t_i, D_2) > \sum_{q_i \in Q} c(q_1, D_1) > 0$, $\sigma'(\{D_1\} - \{Q\}, \{D_2\} - \{Q\}) > \delta$ then $S(Q, D_1) \geq S(Q, D_2)$.

STMC3—Original Let $Q = \{q_1, q_2\}$ be a two-term query and d a non-query term such that $\sigma(d, q_2) > 0$. If $|D_1| = |D_2| > 1$, $c(q_1, D_1) = |D_1|$, $c(q1, D_2) = |D_2| - 1$ and $c(d, D_2) = 1$, then $S(Q, D_1) \leq S(Q, D_2)$.

STMC3—Adapted Let D_1 and D_2 be two arbitrary long documents that covers the same number of query terms (i.e. $|D_1 \bigcap Q| = |D_2 \bigcap Q|$). If $|D_1| - |D_2| \leq abs(\delta)$, $\sum_{q_i \in Q} c(q_i, D_1) > \sum_{q_i \in Q} c(q_i, D_2)$ and $\sigma'(\{D_2\} - \{Q\}, Q) > \sigma'(\{D_1\} - \{Q\}, Q)$, then $S(Q, D_1) > S(Q, D_2)$.

4 Experiments

We create diagnostic datasets for each of these axioms by extracting instances of queries and documents that already exist in the dataset. In this section, we explain how these datasets were generated and how we employed them to evaluate BERT.

Table 1. Overview of retrieval heuristics employed in our work. The diagnostic datasets for heuristics marked with a blue background were first discussed in [25]. The naming of the heuristics is largely taken from the papers proposing them.

Heuristic Instance	**Informal description**
Term frequency constraints	
TFC1 [9]	The more occurrences of a query term a document has, the higher its retrieval score.
TFC2 [9]	The increase in retrieval score of a document gets smaller as the absolute query term frequency increases.
M-TDC [9,30]	The more discriminating query terms (i.e., those with high IDF value) a document contains, the higher its retrieval score.
Length normalization constraints	
LNC1 [9]	The retrieval score of a document decreases as terms not appearing in the query are added.
LNC2 [9]	A document that is duplicated does not have a lower retrieval score than the original document.
Semantic term matching constraints	
STMC1 [11]	A document's retrieval score increases as it contains terms that are more semantically related to the query terms.
STMC2 [11]	The document terms that are a syntactic match to the query terms contribute at least as much to the document's retrieval score as the semantically related terms.
STMC3 [11]	A document's retrieval score increases as it contains more terms that are semantically related to *different* query terms.
Term proximity constraint	
TP [31]	A document's retrieval score increases as the query terms appearing in it appear in closer proximity.

4.1 TREC 2019 Deep Learning Track

In order to extract diagnostic datasets, we used the corpus and queries for the Document Ranking Task from the TREC 2019 Deep Learning track[7]. This is the only publicly available ad-hoc retrieval dataset that was built specifically for the training of deep neural models, with 3,213,835 web documents and 372,206 queries (367,013 queries in the training set and 5,193 in the development set). The queries and documents, while stripped of HTML elements, are not necessarily well-formed as seen in the following examples from the training set:

- `what is a flail chest`
- `a constitution is best described as a(n) __________.`
- `)what was the immediate impact of the success of the manhattan project?`

[7] https://microsoft.github.io/TREC-2019-Deep-Learning/

The queries consist on average of 5.89(±2.51) words while documents consist on average of 1084.88(±2324.22) words.

Most often, one relevant document exists per query (1.04 relevant documents on average). These relevance judgements were made by human judges on a passage-level: if a passage within a document is relevant, the document is considered relevant. Unlike other TREC datasets, like Robust04, there is no topic description or topic narrative.

At the time of this writing, the relevance judgements for the test queries were not available. Therefore, we split the development queries further, in a new *dev* and *test* dataset, in a $70\% - 30\%$ fashion. In the rest of this paper, when we refer to the *test* or *dev* dataset, we are referring to this split. The *train* split remains the same as the original dataset.

4.2 Retrieval and Ranking

We begin by indexing the document collection using the Indri toolkit[8], and retrieve the top-100 results using a traditional retrieval model with just one hyperparameter, namely, QL, with Indri's default setting (Dirichlet smoothing [42] and $\mu = 2500$). Finally, for BERT, we employ Hugging Face's library[9] of DistilBERT [28], a distilled version of the original BERT model, with fewer parameters (66 million instead of 340 million), and thus more efficient to train, but with very similar results.

We fine-tuned our BERT[10] model with 10 negative samples for each positive sample from the training dataset, randomly picked from the top-100 retrieved from QL. We set the maximum input length to 512 tokens. For fine-tuning we used the sequence classification model. It is implemented by adding a fully-connected layer on top of the `[CLS]` token embedding, which is the specific output token of the BERT model that our fine-tuning is based upon.

Given the limitation of BERT regarding the maximum number of tokens, we limited the document length to its first 512 tokens, though we note that alternative approaches exist (e.g., in [40] the BERT scores across a document's passages/sentences are aggregated). In Table 3, we report the retrieval effectiveness in terms of nDCG and MRR for the documents limited to 512 tokens. We rerank the top-100 retrieved documents based on its first 512 tokens. It is clear that BERT is vastly superior to QL with a 40% improvement in $nDCG$ and 25% improvement in MRR.

4.3 Diagnostic Datasets

Given the adapted axioms defined in Section 3, we now proceed on describing how to extract actual datasets from our corpus.

[8] https://www.lemurproject.org/indri.php

[9] https://github.com/huggingface/transformers

[10] Code for fine-tuning DistilBERT and generating the diagnostic datasets is available at https://github.com/ArthurCamara/bert-axioms

Table 2. Overview of the number of instances in each diagnostic dataset (row I), the number of instances within each diagnostic dataset that contain a relevant document (row II) and the fraction of instances among all of row II where the order of the documents according to the axiom is in line with the relevance judgments. LNC2 is based on new documents, thus, it does not have a fraction of agreement

	TFC1	TFC2	M-TDC	LNC1	LNC2	TP	STMC1	STMC2	STMC3
Diagnostic dataset size	119,690	10,682	13,871	14,481,949	7452	3,010,246	319,579	7,321,319	217,104
Instances with a *relevant* document	1,416	17	11	138,399	82	20,559	19,666	70,829	1,626
Fraction of instances agreeing with relevance	0.91	0.29	0.82	0.50	-	0.18	0.44	0.63	0.35

TFC1, TFC2, M-TDC, LNC1 We add tuples of queries and documents $\{q, d_i, d_j\}$ (or $\{q, d_i, d_j, d_k\}$ for TFC2) for every possible pair of documents $\{d_i, d_j\}$ in the top 100 retrieved by QL that follow the assumptions from Section 3, with $\delta = 10$. We also compute IDF for M-TDC on the complete corpus of documents, tokenized by WordPiece [38].

LNC2 We create a new dataset, appending the document to itself $k \in \mathbb{Z}$ times until we reach up to 512 tokens[11]. In contrast to Rennings et al. [25] we only perform this document duplication for our test set, i.e., BERT does not "see" this type of duplication during its training phase. On average, the documents were multiplied $k = 2.6408 \pm 1.603$ times, with a median of $k = 2$.

TP We simply add to our dataset every pair of documents $\{q, d_i, d_j\}$ in the top 100 retrieved documents by QL for a given topic that follow the stated TP assumptions.

STMC1,STMC2, STMC3 We define σ as the cosine distance between the embeddings of the terms and σ' as the cosine distance between the average embeddings. We trained the embeddings using GLoVe [23] on the entire corpus. For STMC3, we set $\delta = 0.2$.

4.4 Results

In Table 2 we list the number of diagnostic instances we created for each diagnostic dataset. In addition, we also performed a sanity check on the extent to which the document order determined by each axiom corresponds to the relevance judgements. Although only a small set of instances from each diagnostic dataset contains a document with a relevant document (row II in Table 2) we already see a trend: apart from TFC1 and M-TDC where 91% and 82% of the diagnostic instances have an agreement between axiomatic ordering and relevance ordering, the remaining axioms are actually not in line with the relevance ordering for most of the instances. This is a first indication that we have to consider an alternative set of axioms, better fit for such a corpus, in future work.

In Table 3 we report the fraction of instances both QL and BERT fulfil for each diagnostic dataset. As expected, QL *correctly* (as per the axiom) ranks the

[11] Note that we only append the document to itself if the final size does not exceed 512.

document pairs or triples most of the time, with the only outlier being TP, where QL performs essentially random—again, not a surprise given that QL is a bag-of-words model. In contrast—with the exception of LNC2, where BERT's ranking is essentially the opposite of what the LNC2 axiom considers correct (with only 6% of the instances ranked correctly)—BERT has not learnt anything that is related to the axioms as the fraction of correctly ranked instances hovers around 50% (which is essentially randomly picking a document order). Despite this lack of axiomatic fulfilment, BERT clearly outperforms QL, indicating that the existing axioms are not suitable to analyse BERT.

The reverse ranking BERT proposes for nearly all of the LNC2 instances can be explained by the way the axiom is phrased. It is designed to avoid over penalising documents, and thus a duplicated document should always have a retrieval score that is not lower than the original document. The opposite argument though could be made too (and BERT ranks accordingly), that a duplicated document should not yield a higher score than the original document as it does not contain novel/additional information. As we did not provide LNC2 instances in the training set, BERT is not able to rank according to the axiom, in line with the findings of other neural approaches as shown by Rennings et al. [25].

Finally, we observe that, counter-intuitively, BERT does not show a performance better than QL for semantic term matching constraints. For instance, one may expect that BERT would fare quite well on STMC1, given its semantic nature. However, our results indicate that BERT is actually considers term matching as one of its key features. In order to further explore this tension between semantic and syntactic term matching, we split the queries in our test set by their term overlap between the query and the relevant document (if several relevant documents exist for a query, we randomly picked one of them). If a query/document has no (or little) term overlap, we consider this as a semantic match.

Table 3. Overview of the retrieval effectiveness (nDCG columns) and the fraction of diagnostic dataset instances each model ranks correctly.

	nDCG	MRR	TFC1	TFC2	M-TDC	LNC1	LNC2	TP	STMC1	STMC2	STMC3
QL	0.2627	0.3633	**0.99**	**0.70**	**0.88**	**0.50**	**1.00**	0.39	0.49	**0.70**	**0.70**
DistilBERT	**0.3633**	**0.4537**	0.61	0.39	0.51	**0.50**	0.00	**0.41**	**0.50**	0.51	0.51

The results of this query split can be found in Figure 1. We split the query set roughly into three equally sized parts based on the fraction of query terms appearing in the relevant document (as an example, if a query/document pair has a fraction of 0.5, half of all query terms appear in the document). We report results for all queries (Figure 1 (left)), as well as only those where the relevant document appears in the top-100 QL ranking (Figure 1 (right)). We find that BERT outperforms QL across all three splits, indicating that BERT is indeed able to pick up the importance of syntactic term matching. At the same time, as expected, BERT is performing significantly better than QL for queries that require a large amount of semantic matching. That brings into question on why, then, the axiomatic performance across our semantic axiomatic datasets does not reflect that. One hypothesis is that the semantic similarity we measure (based on context-free word embeddings) is different to the semantic similarity measured via contextual word embeddings. This in itself is an interesting avenue for future work, since it brings a new question on *what* that semantic relationship is, and how to accurately measure it.

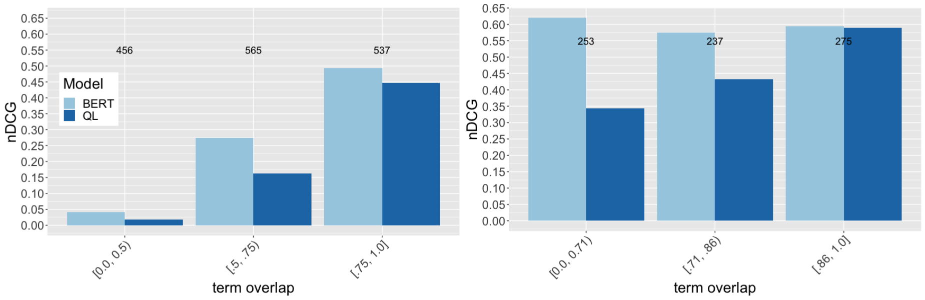

Fig. 1. The test queries are split into three sets, depending on the fraction of term overlap between the query and its corresponding relevant document. On the left, we plot all queries, on the right only those queries for which the relevant document appears in the top-100 ranked documents of the QL ranking.

5 Discussion & Conclusion

In this paper, we set out to analyze BERT with the help of the recently proposed *diagnostic datasets for IR based on retrieval heuristics* approach [25]. We expected BERT to perform better at fulfilling some of the proposed semantic axioms. Instead, we have shown that BERT, while significantly better than traditional models for ad-hoc retrieval, does not fulfil most retrieval heuristics, created by IR experts, that are supposed to produce better results for ad-hoc retrieval models. We argue that based on these results, the axioms are not suitable to analyse BERT and it is an open question what type of axioms would be able to capture some performance aspects of BERT and related models. In fact, how to arrive at those additional axioms, based on the knowledge we have now gained about BERT is in itself an open question.

Acknowledgement. This research has been supported by NWO project SearchX (639.022.722).

References

1. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP". ACL (2018)

2. Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C.: UMass at TREC 2004: Novelty and HARD. Computer Science Department Faculty Publication Series p. 189 (2004)
3. Borlund, P.: The Concept of Relevance in IR. Journal of the American Society for information Science and Technology **54**(10), 913–925 (2003)
4. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What Does BERT Look At? An Analysis of BERT's Attention. CoRR **abs/1906.04341** (2019)
5. Clinchant, S., Gaussier, E.: Is Document Frequency Important for PRF? In: Conference on the Theory of Information Retrieval. pp. 89–100. Springer (2011)
6. Dai, Z., Callan, J.: Deeper Text Understanding for IR with Contextual Neural Language Modeling. In: SIGIR. pp. 985–988. ACM (2019)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT (1). pp. 4171–4186. ACL (2019)
8. Fang, H.: A Re-examination of Query Expansion Using Lexical Resources. In: ACL. pp. 139–147. ACL (2008)
9. Fang, H., Tao, T., Zhai, C.: A Formal Study of Information Retrieval Heuristics. In: SIGIR. pp. 49–56. ACM (2004)
10. Fang, H., Zhai, C.: An Exploration of Axiomatic Approaches to Information Retrieval. In: SIGIR. pp. 480–487. ACM (2005)
11. Fang, H., Zhai, C.: Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In: SIGIR. pp. 115–122. ACM (2006)
12. Fernando, Z.T., Singh, J., Anand, A.: A Study on the Interpretability of Neural Retrieval Models using DeepSHAP. In: SIGIR. pp. 1005–1008. ACM (2019)
13. Hagen, M., Völske, M., Göring, S., Stein, B.: Axiomatic Result Re-Ranking. In: CIKM. pp. 721–730. ACM (2016)
14. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional Neural Network Architectures for Matching Natural Language Sentences. In: NIPS. pp. 2042–2050 (2014)
15. Jia, R., Liang, P.: Adversarial Examples for Evaluating Reading Comprehension Systems. In: EMNLP. pp. 2021–2031. ACL (2017)
16. Lavrenko, V., Croft, W.B.: Relevance-Based Language Models. In: SIGIR. pp. 120–127. ACM (2001)
17. Lin, J.: The Neural Hype and Comparisons Against Weak Baselines. SIGIR Forum **52**(2), 40–51 (2018)
18. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: CEDR: contextualized embeddings for document ranking. In: SIGIR. pp. 1101–1104. ACM (2019)
19. McDonald, R., Brokos, G., Androutsopoulos, I.: Deep Relevance Ranking using Enhanced Document-Query Interactions. In: EMNLP. pp. 1849–1860. ACL (2018)
20. Mitra, B., Craswell, N.: An Introduction to Neural Information Retrieval. Foundations and Trends in Information Retrieval **13**(1), 1–126 (2018)
21. Niven, T., Kao, H.: Probing Neural Network Comprehension of Natural Language Arguments. In: ACL (1). pp. 4658–4664. ACL (2019)
22. Pang, L., Lan, Y., Guo, J., Xu, J., Cheng, X.: A Deep Investigation of Deep IR Models. CoRR **abs/1707.07700** (2017)
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: EMNLP. pp. 1532–1543. ACL (2014)
24. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners (2019)
25. Rennings, D., Moraes, F., Hauff, C.: An Axiomatic Approach to Diagnosing Neural IR Models. In: ECIR (1). Lecture Notes in Computer Science, vol. 11437, pp. 489–503. Springer (2019)

26. Robertson, S.E., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (2009)
27. Rosset, C., Mitra, B., Xiong, C., Craswell, N., Song, X., Tiwary, S.: An Axiomatic Approach to Regularizing Neural Ranking Models. In: SIGIR. pp. 981–984. ACM (2019)
28. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter (2019)
29. Saracevic, T.: Relevance Reconsidered. In: CoLIS 2. pp. 201–218. ACM (1996)
30. Shi, S., Wen, J.R., Yu, Q., Song, R., Ma, W.Y.: Gravitation-based Model for Information Retrieval. In: SIGIR. pp. 488–495. ACM (2005)
31. Tao, T., Zhai, C.: An Exploration of Proximity Measures in Information Retrieval. In: SIGIR. pp. 295–302. ACM (2007)
32. Tenney, I., Das, D., Pavlick, E.: BERT Rediscovers the Classical NLP Pipeline. In: ACL (1). pp. 4593–4601. ACL (2019)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. In: NIPS. pp. 5998–6008 (2017)
34. Verma, M., Ganguly, D.: LIRME: Locally Interpretable Ranking Model Explanation. In: SIGIR. pp. 1281–1284. ACM (2019)
35. Voorhees, E.M.: Overview of the TREC 2004 Robust Retrieval Track. In: TREC (2004)
36. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: BlackboxNLP@EMNLP. pp. 353–355. ACL (2018)
37. Weston, J., Bordes, A., Chopra, S., Mikolov, T.: Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In: ICLR (Poster) (2016)
38. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. CoRR **abs/1609.08144** (2016)
39. Yang, L., Ai, Q., Guo, J., Croft, W.B.: aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model. In: CIKM. pp. 287–296. ACM (2016)
40. Yang, W., Zhang, H., Lin, J.: Simple Applications of BERT for Ad Hoc Document Retrieval. CoRR **abs/1903.10972** (2019)
41. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: NeurIPS. pp. 5754–5764 (2019)
42. Zhai, C., Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Ad hoc Information Retrieval. In: ACM SIGIR Forum. vol. 51, pp. 268–276. ACM (2017)