An Axiomatic Approach to Diagnosing Neural IR Models

Daniël Rennings[1], Felipe Moraes[2], and Claudia Hauff[2]

[1] d.j.a.rennings@student.tudelft.nl
[2] {f.moraes,c.hauff}@tudelft.nl
Delft University of Technology, Delft, the Netherlands

Abstract. Traditional retrieval models such as BM25 or language models have been engineered based on search heuristics that later have been formalized into axioms. The axiomatic approach to information retrieval (IR) has shown that the effectiveness of a retrieval method is connected to its fulfillment of axioms. This approach enabled researchers to identify shortcomings in existing approaches and "fix" them. With the new wave of neural net based approaches to IR, a theoretical analysis of those retrieval models is no longer feasible, as they potentially contain millions of parameters. In this paper, we propose a pipeline to create *diagnostic datasets for IR*, each engineered to fulfill one axiom. We execute our pipeline on the recently released large-scale question answering dataset `WikiPassageQA` (which contains over 4000 topics) and create diagnostic datasets for four axioms. We empirically validate to what extent well-known deep IR models are able to realize the axiomatic pattern underlying the datasets. Our evaluation shows that there is indeed a positive relation between the performance of neural approaches on diagnostic datasets and their retrieval effectiveness. Based on these findings, we argue that diagnostic datasets grounded in axioms are a good approach to diagnosing neural IR models.

1 Introduction

Over the past few years, deep learning approaches have been increasingly applied to IR tasks [28]. At the same time, the IR community has identified a number of issues [7,8,28], hindering the kind of progress seen in other research areas such as natural language processing (NLP) and computer vision. Overall, our community lacks adequately large-scale public datasets for training (an exception is the recently released `WikiPassageQA` [6]), shared public code repositories of neural IR models (although some progress has been recently made [11,21]) and approaches to interpret and analyze neural IR models (here, [5] is an exception).

In this paper, we focus our attention on the last issue—the analysis of neural IR models. While many neural models have been proposed, few have turned out to outperform properly tuned BM25 or language modeling baselines. Traditional retrieval models have been engineered based on search heuristics that later have been formalized into *axioms*—formal constraints that should be fulfilled by a good model—which enable us to analytically investigate to what

extent retrieval models fulfill them [13,15,16,14]. This analytical approach enabled researchers to identify shortcomings in existing retrieval models and "fix" them [18,1,12,27,4,30], in order to achieve higher retrieval effectiveness. Ideally, we employ a similar axiomatic approach to diagnose & fix neural IR models in order to reap the benefits deep learning has offered in other fields. However, as these models may contain millions of parameters [28], this is not possible.

Instead, we propose a pipeline for the creation of *diagnostic datasets for IR*, each engineered to fulfill one axiom. This approach follows the tradition in NLP and computer vision where dataset creation for diagnostic purposes is well-known—consider for instance `bAbI` [44] for automatic text understanding & reasoning, adversarial examples for `SQUAD` [22], a popular reading comprehension dataset, and `CLEVR` [23], a dataset for language & visual reasoning.

We execute our pipeline for four axioms on the answer-passage retrieval dataset `WikiPassageQA` [6] which contains more than 4,000 topics. It has been shown to be a difficult dataset for a range of neural models. We empirically validate to what extent four well-known deep IR models are able to realize the axiomatic patterns underlying the datasets. We find that, indeed, there is a positive relation between the performance of neural approaches on the diagnostic datasets and their retrieval effectiveness.

We believe these findings to be more insightful for IR researchers to improve neural models than, for instance, the probing of neural net layers via NLP tasks [5] or simply evaluating deep models with a range of metrics on standard test collections. The **main contribution** of our work is to showcase that a transformation from an analytical axiom to a diagnostic dataset is possible and offers us a new tool to diagnose retrieval models that are too complex to be analyzed theoretically.

2 Related Work

Axiomatic approach to IR Fang et al.'s [13] seminal work introduced six retrieval constraints (later coined *axioms* [15]) that a reasonable retrieval function should satisfy. Formalizing retrieval heuristics into constraints—e.g. *given a single-term query w and two equally long documents, the retrieval score of the document with a higher frequency of w should be ranked higher* (also known as constraint `TFC1`)—enabled the authors to *analytically* evaluate a number of existing retrieval functions. The main assumption of the work—the effectiveness of retrieval functions is connected to their fulfillment of retrieval constraints—was empirically validated. Fang et al. [14] also proposed the use of perturbed document collections to gather further insights on retrieval functions fulfilling the same set of axioms. This approach has not been followed-up upon in works other than [31].

Apart from the diagnosis of existing functions, Fang et al. derived novel retrieval functions based on their initial set of constraints [15] and later extended their list of axioms from purely term-matching to semantic-matching based constraints [16,12]. Others have contributed query term proximity [42,18], document

length normalization [27] and query term discrimination [1] constraints, consistently showing that traditional models improve when slightly altered to satisfy those constraints. While most of the more than twenty existing axioms have been designed for standard retrieval models, a number of axioms have been proposed for more specialized cases, such as statistical translation models [24] and pseudo-relevance feedback [3,4,30].

Lastly, we point to two works closest to ours. Hagen et al. [18] explored the re-ranking of a given result list based on the aggregated re-ranking preferences of twenty-three axioms. Similar to our work, this application of axioms to an actual result list (instead of "hypothetical" documents containing one or two different query terms used in the analytic evaluation of retrieval functions) requires the *extension* and *relaxation* of axioms. Pang et al. [35] investigated differences in neural IR models and learning to rank approaches with hand-crafted features. Through a manual error analysis, weaknesses in deep IR models were identified and connected to retrieval constraints.

Neural IR models By now, deep learning has become the mainstay in a number of research fields, yielding impressive improvements on long-standing tasks. The information retrieval community has also seen a large number of proposed neural IR models, which can be categorized as *interaction-based*, *representation-based* or a *hybrid* between the two [17], based on the manner they model the query and document. While interaction-based neural approaches (e.g., DeepMatch [26], DRMM [17], MatchPyramid [36], ANMM [45]) use the local interactions between the query and document as input to the deep net, representation-based approaches (e.g., DSSM [20], C-DSSM [39], ARC-I [19]) strive to create good representations of the query and the document separately; hybrid approaches such as Duet [29], ARC-II [19] and MVLSTM [43] incorporate both an interaction- and representation-based component. Despite the motivation for representation-based approaches and the need for semantics over syntax matching, a recent comparative study [32] has shown the deep interaction-based approaches to clearly outperform the representation-based approaches in terms of retrieval effectiveness. Whereas most interaction- and representation-based approaches compute relevance at the document level, Fan et al. [10] recently proposed a hierarchical neural matching model (HiNT) which employs a local matching layer and global decision layer, to capture relevance signals at the passage and document level which compete with each other. Another recent work has achieved state-of-the-art performance by creating a neural pseudo relevance feedback framework (NPRF) that can be used with existing neural IR models as building blocks [25].

While many works have presented novel neural approaches, few works have focused on diagnosing neural IR models. While studies such as the one conducted in [32] enable us to empirically determine which type of approach performs better, they can only provide relatively coarse-grained insights (in this case: interaction-based performs better than representation-based). In contrast, Cohen et al. [5] recently proposed to *probe* neural retrieval models by training them,

and then using each layer's weights as input to a classifier for different types of NLP tasks (sentiment analysis, POS tagging, etc). The performance on those tasks provides insights into the kind of information that each layer captures. While this is indeed useful to realize, it does not provide an immediate insight into how to improve an existing neural approach. It is also quite labor-intensive as this probing is not model agnostic—in contrast to our work.

While in the IR community, the diagnosing of deep nets is in its infancy, the computer vision and NLP communities have proposed a number of different manners to open up this black box that a typical deep net is. The 20 `bAbI` tasks [44] were developed specifically to diagnose text understanding and reasoning systems, while Jia and Liang [22] proposed an adversarial evaluation scheme of the `SQUAD` dataset by inserting distracting sentences into text passages (and as a result all evaluated models dropped sharply in their accuracy). `CLEVR` [23], a dataset for language and visual reasoning, consists of a large number of rendered images (constructed from a limited universe of objects and relationships) and automatically generated questions.

Here, we propose to bring the approach of diagnostic dataset creation into the IR community, based on well-established axioms.

3 Creating Diagnostic Datasets

Out of the more than twenty IR axioms proposed by now, we have selected four among those in [13,40,14], and converted them for our purpose of diagnostic dataset creation. Two of the axioms (`TFC1` and `M-TDC`) were selected as they capture a fair amount of relevance, while being present in existing datasets—including the one we work with. Combined, `TFC1` and `M-TDC` essentially represent the TF-IDF statistic, a pervasive component in most IR models [48,49,6,9]. A third axiom (`TFC2`) constraints the difference in scores between pairs of documents instead of individual documents. We include `TFC2` to show our methodology can handle such axioms as well. Finally, we selected the `LNC2` axiom to showcase how we can generate a diagnostic dataset from an existing corpus through creating artificial data when extracting a diagnostic dataset does not yield enough data points.

We now describe the axioms we consider and propose (1) an *extension* of each axiom in order to match realistic queries and documents[3], and, (2) a *relaxation* of extended axioms such that the strictly defined query and document relations are relaxed to enable selection and generation of sufficient amounts of data. Whereas step (1) allows us to move from one- or two-term queries to arbitrary query lengths and from two- or three-document instances to any number of documents, step (2) allows us to make use of query/document pairings that *approximately* fulfill a particular relationship.

[3] For completeness, we note that we did not observe a single instance of query-document pairs or triplets in our `WikiPassageQA` corpus that satisfies any of the four original (non-extended, non-relaxed) axioms considered here.

Finally, a note on notation: we refer to an original axiom as `Axiom`; its extended and relaxed variant is referred to as $\overline{\texttt{Axiom}}$.

3.1 TFC1: extension and relaxation

The TFC1 axiom [13] favours documents with more occurrences of a query term and is formally defined as follows: let $\mathbf{q} = \{w\}$ be a single-term query and $\mathbf{d_1}$ and $\mathbf{d_2}$ be two documents of equal length, i.e. $|\mathbf{d_1}| = |\mathbf{d_2}|$. Further, let $c(w, \mathbf{d})$ be the count of term w in document $\mathbf{d}$ and $S(\mathbf{q}, \mathbf{d})$ be the retrieval status value a retrieval function assigns to $\mathbf{d}$, given $\mathbf{q}$. TFC1 then states that if $c(w, \mathbf{d_1}) > c(w, \mathbf{d_2})$ holds, $S(\mathbf{q}, \mathbf{d_1}) > S(\mathbf{q}, \mathbf{d_2})$ should hold as well.

We now *extend* this axiom to multiple-term queries and *relax* it to incorporate documents of approximately the same length, resulting in $\overline{\texttt{TFC1}}$. Formally: let $\mathbf{q} = \{w_1, w_2, ..w_{|\mathbf{q}|}\}$ be a multi-term query and $|\mathbf{d_i}| \approx |\mathbf{d_j}|$, i.e. $|\mathbf{d_i}| - |\mathbf{d_j}| \leq |\delta_{\texttt{TFC1}}|$. Here, $\delta_{\texttt{TFC1}}$ is an adjustable parameter that may be set according to the document corpus and retrieval task. Additionally, we *relax* the constraint that $\mathbf{d_i}$ must have a larger count for *every* query term than $\mathbf{d_j}$. We now require $c(w, \mathbf{d_i}) \geq c(w, \mathbf{d_j})\ \forall w \in \mathbf{q}$ and $\sum_{w \in \mathbf{q}} c(w, \mathbf{d_i}) > \sum_{w \in \mathbf{q}} c(w, \mathbf{d_j})$, i.e. there is at least one query term with a higher term count in $\mathbf{d_i}$. If this relaxed constraint is fulfilled, then $\overline{\texttt{TFC1}}$ states that $S(\mathbf{q}, \mathbf{d_i}) > S(\mathbf{q}, \mathbf{d_j})$.

3.2 TFC2: extension and relaxation

Axiom TFC2 [13] encapsulates the intuition that an increase in retrieval status value due to an increase in term count becomes smaller as the absolute term count increases. Formally, the axiom considers the case of $\mathbf{q} = \{w\}$ and $|\mathbf{d_1}| = |\mathbf{d_2}| = |\mathbf{d_3}|$. If $c(w, \mathbf{d_1}) > 0$, $c(w, \mathbf{d_2}) - c(w, \mathbf{d_1}) = 1$ and $c(w, \mathbf{d_3}) - c(w, \mathbf{d_2}) = 1$ (i.e. the absolute term count of w in $\mathbf{d_1}$ is smallest and in $\mathbf{d_3}$ is largest), then $S(\mathbf{q}, \mathbf{d_2}) - S(\mathbf{q}, \mathbf{d_1}) > S(\mathbf{q}, \mathbf{d_3}) - S(\mathbf{q}, \mathbf{d_2})$.

We define $\overline{\texttt{TFC2}}$ for multi-term queries and documents of approximately the same length. Formally, we consider $\mathbf{q} = \{w_1, w_2, ..w_{|\mathbf{q}|}\}$ and $|\mathbf{d_i}| \approx |\mathbf{d_j}| \approx |\mathbf{d_k}|$, i.e. $max_{\mathbf{d_a}, \mathbf{d_b} \in \{\mathbf{d_i}, \mathbf{d_j}, \mathbf{d_k}\}}(|\mathbf{d_a}| - |\mathbf{d_b}|) \leq |\delta_{\texttt{TFC2}}|$. Every document has to contain at least one query term and the differences in term count are no longer restricted to be exactly 1. This leads to the constraints $\sum_{w \in \mathbf{q}} c(w, \mathbf{d_k}) > \sum_{w \in \mathbf{q}} c(w, \mathbf{d_j}) > \sum_{w \in \mathbf{q}} c(w, \mathbf{d_i}) > 0$ and $c(w, \mathbf{d_j}) - c(w, \mathbf{d_i}) = c(w, \mathbf{d_k}) - c(w, \mathbf{d_j})\ \forall w \in \mathbf{q}$. The latter constraint does not mean that the difference has to be the same for every query term, instead we enforce this equality in term count difference on a term level. If these constraints hold, then according to $\overline{\texttt{TFC2}}$, $S(\mathbf{q}, \mathbf{d_j}) - S(\mathbf{q}, \mathbf{d_i}) > S(\mathbf{q}, \mathbf{d_k}) - S(\mathbf{q}, \mathbf{d_j})$.

3.3 M-TDC: extension and relaxation

The TDC axiom was originally proposed by Fang et al. [13] to favour documents with more occurrences of less popular query terms in the collection. Shi et al. modified the TDC axiom to M-TDC [40] to fix undesired behavior in some cases.

Formally, M-TDC is defined as follows. Let $\mathbf{q} = \{w_1, w_2\}$ be a two-term query and assume $|\mathbf{d_1}| = |\mathbf{d_2}|$, $c(w_1, \mathbf{d_1}) = c(w_2, \mathbf{d_2})$ and $c(w_2, \mathbf{d_1}) = c(w_1, \mathbf{d_2})$. If $idf(w_1) \geq idf(w_2)$—i.e. w_1 is rarer in the corpus than w_2—and $c(w_1, \mathbf{d_1}) \geq c(w_1, \mathbf{d_2})$, then $S(q, \mathbf{d_1}) \geq S(q, \mathbf{d_2})$.

We define $\overline{\texttt{M-TDC}}$ for multi-term queries $\mathbf{q} = \{w_1, w_2, ..w_{|\mathbf{q}|}\}$ and pairs of documents $\mathbf{d_i}, \mathbf{d_j}$ that (i) differ in at least one count of a query term, (ii) have the same total count of query terms (sum of term frequencies), and, (iii) have approximately the same length, i.e. $|\mathbf{d_i}| - |\mathbf{d_j}| \leq |\delta_{\texttt{M-TDC}}|$.

For a query-doc-doc triplet to be included in our axiomatic dataset with retrieval score preference $S(\mathbf{q}, \mathbf{d_i}) \geq S(\mathbf{q}, \mathbf{d_j})$, all query terms for which $c(w, d_i) \neq c(w, d_j)$ need to appear at least once in a *valid* query term pair. We evaluate all possible query term pairs (with $w_a \neq w_b$) and consider a query term pair to be *valid* when the following conditions hold: (i) $idf(w_a) \geq idf(w_b)$, (ii) $c(w_a, \mathbf{d_i}) = c(w_b, \mathbf{d_j})$ and $c(w_b, \mathbf{d_i}) = c(w_a, \mathbf{d_j})$, (iii) $c(w_a, \mathbf{d_i}) > c(w_a, \mathbf{d_j})$, and (iv) $c(w_a, \mathbf{q}) \geq c(w_b, \mathbf{q})$.

3.4 LNC2: extension and relaxation

The LNC2 [14] axiom prescribes that over-penalizing long documents should be avoided: if a document is replicated k times, its retrieval status value should not be lower than that of its un-replicated variant. The axiom was defined under the assumption that redundancy is not an issue, which we also follow here. Formally the axiom is defined as follows: let $\mathbf{q} = \{w_q\}$, $c(w_q, \mathbf{d_1}) > 0$, $k > 1, k \in \mathbb{N}$, $|\mathbf{d_1}| = k \times |\mathbf{d_2}|$, and for $\forall w \in \mathbf{d_1}$, $c(w, \mathbf{d_1}) = k \times c(w, \mathbf{d_2})$. If those constraints are met, $S(\mathbf{q}, \mathbf{d_1}) \geq S(\mathbf{q}, \mathbf{d_2})$ should hold. This axiom can simply be extended to $\overline{\texttt{LNC2}}$ by defining $\mathbf{q}$ for multi-term queries and documents $\mathbf{d_i}$ and $\mathbf{d_j}$. No additional relaxation is required.

3.5 From $\overline{\texttt{Axiom}}$ to dataset

Having defined extended and relaxed variants of our axioms, we now describe how to obtain a diagnostic dataset for each. Given a corpus with standard pre-processing applied, we determine the number of instances the (i) original axiom, (ii) relaxed axiom and (iii) relaxed & extended axiom can be found in it. As the axioms are defined over retrieval status values (instead of relevance labels), we do not require relevance judgments and thus, almost any dataset is suitable as source dataset. We can sample queries and document pairs/triplets from such a dataset at will; we keep those in our diagnostic datasets that satisfy our axioms. Due to the very restrictive nature of the original axioms, we expect few instances that fulfill their conditions to be found in most existing corpora. In the case of the extended axiom $\overline{\texttt{LNC2}}$, we expect only spam documents to satisfy the axiom. For this axiom we move beyond *extracting* instances from a given corpus and artificially *create* instances instead by appending each selected document in the dataset $k - 1$ times to itself for a set of values $k > 1$. Figure 1 shows a graphical overview of our pipeline, with examples from WikiPassageQA.

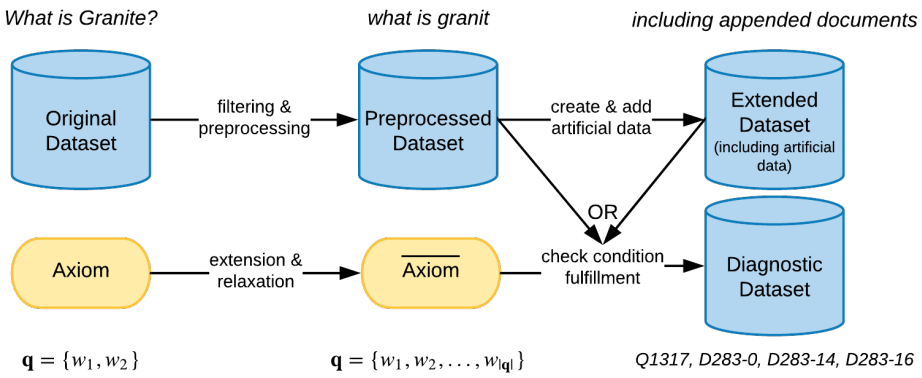

Fig. 1: Overview of the diagnostic dataset creation pipeline. In *italics*, we show an example for the $\overline{\texttt{TFC2}}$ axiom as extracted from question 1317 on passages from Wikipedia document 283 in the `WikiPassageQA` dataset, and refer to appended documents as an example of artificial data (for $\overline{\texttt{LNC2}}$).

3.6 Evaluating IR models

The evaluation through diagnostic datasets as presented in this work is **model-agnostic**. Given a trained and tuned model, we record the fraction of diagnostic dataset instances that the model scores according to the axiomatic preferences. Thus, a model that is able to rank all instances correctly achieves an *axiomatic score* of 1.0. In line with past works on axiomatic approaches to IR, we expect there to be a positive correlation between models' retrieval effectiveness and their axiomatic scores.

4 Experiments

We now introduce the corpus we use for our study in more detail, then discuss details of the diagnostic datasets created from this corpus. Subsequently, we introduce the employed retrieval models and finally explore to what extent our traditional baselines and neural IR models satisfy the constraints encapsulated in the diagnostic datasets and how this relates to their retrieval effectiveness.

4.1 `WikiPassageQA`

We empirically validate our diagnostic dataset creation pipeline on the answer passage retrieval dataset `WikiPassageQA` [6]. As it contains thousands of topics, it is a suitably large dataset for the training of neural models.

The dataset consists of 861 Wikipedia documents, split into passages of six sentences each, yielding 50,477 unique passages in total—each containing 135.2 words on average. Crowd-workers created a total of 4,154 questions. For example, for the Wikipedia document on *Granite*[4] the created questions include:

[4] https://en.wikipedia.org/wiki/Granite

- *What is the occurrence of granite?*
- *How does weathering affect granite?*
- *What are the geochemical origins of granite?*

The questions contain 9.5 terms on average (minimum 2[5], maximum 39). The binary passage-level relevance judgments were also sourced from the same crowdworkers and later validated by a subsequent mechanical turk verification poll. On average, there are 1.7 relevant passages per question[6].

The corpus has been developed for the *answer passage retrieval* task: given a query (the question) and a Wikipedia document (more concretely, all passages making up that document), rank the passages such that those containing the answer to the question are ranked on top. `WikiPassageQA` has been released with a pre-defined train/dev/test split that we maintain in our work.

As in [6], we employ mean average precision (MAP), mean reciprocal rank (MRR) and precision at k documents (P@k) to report retrieval effectiveness. As noted earlier, in terms of *axiomatic performance*, we report the fraction of precedence constraints each model satisfies.

In terms of pre-processing, we apply stemming[7] but not stopword removal, as the latter may actually remove informative terms from the text such as the question words *what* and *why*.

4.2 Diagnostic datasets

Given the nature of our corpus, we do not need to randomly sample document pairs or triplets. Instead, we consider all possible pairs or triplets (depending on the axiom in question) of passages within a single Wikipedia document and the respective questions and keep those instances in our diagnostic datasets that satisfy our extended and relaxed axiomatic constraints—once more, keeping in place the train/dev/test split of the original corpus. Since the Wikipedia documents are already split into six-sentence passages, we do not manually set a threshold (δ) of allowed document length differences for this corpus and instead accept all passage pairs/triplets as sufficiently similar in length.

For the $\overline{\texttt{LNC2}}$ axiom, we have two options: we can either (1) add duplicated documents to the test set only (which may be considered "unfair" to the neural models as they have never seen this type of document in the training data) or (2) we add duplicated documents to all (train/dev/test) splits. Here, we report the axiomatic performance of our investigated models across both options.

In Table 1 we report the number of extracted diagnostic instances per axiom. Let us first consider the three axioms ($\overline{\texttt{TFC1}}$, $\overline{\texttt{TFC2}}$ and $\overline{\texttt{M-TDC}}$) based on data extraction: depending on the axiom, we extract between 42K and 3.5M instances.

[5] The question is *define Hydroelectricity*.

[6] The corpus statistics reported here differ slightly from those reported in [6] as we filtered out instances with empty question texts, duplicated questions and questions appearing in both the training and test set.

[7] We employed the `nltk.stem.SnowballStemmer` for the English language.

One may question the need for the introduced relaxations that go beyond document length relaxation. We incorporated those as document length relaxation alone was insufficient for this corpus: as a concrete example, for $\overline{\texttt{TFC1}}$ (when extended and *only* with document length relaxation applied), we only found six instances that could be extracted from `WikiPassageQA`.

Let us now consider $\overline{\texttt{LNC2}}$, whose instances are not extracted from the corpus, but instead require data generation based on the original corpus. We created instances with $k = \{2, 3, 4\}$ times the original content and maintain the original labels (e.g. a passage that was labelled relevant in its original form is labelled relevant in its artificial form as well, as supported by the `LNC2` axiom). We only considered passages up to 240 words in eventual length, due to experimental constraints[8], leading to a total of 10K and 100K instances respectively for the two variants of $\overline{\texttt{LNC2}}$. Note, that for $\overline{\texttt{LNC2}}^{All}$ we train the neural models not only on the original train split of `WikiPassageQA`, but add the generated instances to the training data as well; this addition does not significantly alter the fraction of relevant to non-relevant answer passages.

Table 1: Number of instances per axiom ($\overline{\texttt{TFC1}}$, $\overline{\texttt{TFC2}}$, $\overline{\texttt{M-TDC}}$) extracted from `WikiPassageQA`. For $\overline{\texttt{LNC2}}$ we report the number of artificial diagnostic instances, in two variants: the duplication of document content restricted to the test set ($\overline{\texttt{LNC2}}^{Test}$) and across the train/dev/test sets ($\overline{\texttt{LNC2}}^{All}$).

	$\overline{\texttt{TFC1}}$	$\overline{\texttt{TFC2}}$	$\overline{\texttt{M-TDC}}$	$\overline{\texttt{LNC2}}^{Test}$	$\overline{\texttt{LNC2}}^{All}$
Parameters				$k = \{2, 3, 4\}$, $doc\_len_{max} = 240$	
Train	2,758,223	837,838	33,509	0	82,785
Dev	376,902	50,772	3,958	0	10,485
Test	353,621	183,898	4,497	10,074	10,074
Total	3,488,746	1,072,508	41,964	10,074	103,344

4.3 Retrieval models

For our experiments we opted for the retrieval baselines BM25 and query likelihood with Dirichlet smoothing (QL) as implemented in the `Indri` toolkit [41] (version 5.11) and four neural IR models as implemented in `MatchZoo`[9] [11].

We tuned the hyper-parameters of BM25 and QL on the train and development parts of `WikiPassageQA`, optimizing for MAP, resulting in the following settings: $k1 = 0.4, b = 0.1, k3 = 1$ (BM25) and $\mu = 750$ (QL).

For our neural models, we employed the `MatchZoo` retrieval toolkit which has been employed in a number of prior studies, including [38,47,2]. `MatchZoo`

[8] Concretely, we use the `MatchZoo` toolkit for our neural models and ran into issues when the maximum document length was set to include longer passages, see also https://github.com/faneshion/MatchZoo/issues/264.

[9] Version https://github.com/NTMC-Community/MatchZoo/tree/e564565.

contains architecture configurations[10] that have been optimized for the `WikiQA` dataset [46], an open-domain question answering dataset, similar in spirit to `WikiPassageQA`, though defined on the document, not passage level. Due to the computational requirements of neural model training, we limited the maximum query length and passage length to 20 and 240 terms respectively—as a result, in more than 99% of all training instances the entire question and entire passage was considered. We maintained the default `MatchZoo` configurations, including learning rates and optimizers. All neural models were trained for 400 iterations.

Initially, we considered all neural models implemented in `MatchZoo`; however, for a number of models (especially the models incorporating a representation-based component such as CDSSM and MV-LSTM) we observed a significant drop in retrieval effectiveness in the `WikiPassageQA` dataset compared to `WikiQA` when relying on the pre-configured model architectures. As a concrete example, MV-LSTM dropped in MAP from 0.62 in `WikiQA` to 0.22 in `WikiPassageQA`. This lack of model robustness to the corpus is a well-known problem for neural models. Since neural architecture search [50] is beyond the scope of our work, we here consider the four best-performing models, which are all interaction-based, in line with the findings reported in [32]. Concretely, the four models are:

- DRMM [17], an interaction-based model that employs a histogram representation of the similarity between a query and a document;
- aNMM [45], an attention based neural matching model, specifically designed for ranking short text in an interaction-based fashion;
- Duet [29], a hybrid of an interaction-based and representation-based model: it combines two separate deep neural networks, one employs a local representation, and another employs distributed representations for matching the query and the document;
- MatchPyramid [36], a hybrid model that mimics image recognition in its text matching and employs a convolutional neural network.

4.4 Retrieval model performance

The main results of our study are shown in Table 2 where we present the models' retrieval effectiveness on the original `WikiPassageQA` corpus as well as the fraction of axiomatic instances each model ranks correctly.

Let us first consider the retrieval effectiveness of our models. As found in several prior studies [37,34,17], and as already indicated in [6] with regard to the `WikiPassageQA` dataset, neural models struggle to outperform decades-old retrieval baselines that contain just a handful of hyper-parameters (and recall that we only report the best-performing neural models wrt. retrieval effectiveness). Only DRMM and aNMM are able to significantly outperform the traditional models, with an increase in MAP from 0.54 (QL) to 0.55 (DRMM) and 0.57 (aNMM) respectively. These results are not unexpected, as DRMM is considered to be one of the most competitive neural IR models to date [33], and DRMM

[10] https://github.com/faneshion/MatchZoo/tree/e564565/examples/wikiqa/config contains the configurations (learning rate, optimizer, etc.) per model.

and aNMM are similar in the sense that they both model the interaction between query terms and document terms to build a matching matrix [33]. Furthermore, similar to [37] we find that DRMM outperforms MatchPyramid, and similar to [21] we observe that MatchPyramid in turn outperforms Duet.

Moving on to the axiomatic performance of our models, we find both BM25 and QL to satisfy the precedence constraints of the vast majority of instances across all four axioms: both models satisfy more than 90% of the $\overline{\texttt{M-TDC}}$ instances and more than 70% of the $\overline{\texttt{TFC1}}$ instances. The largest difference in percentage of satisfied axiomatic instances can be found in $\overline{\texttt{TFC2}}$ (BM25 satisfies 98% of instances, QL only 63%), which can explained by the fact that QL with Dirichlet smoothing employs a document length dependent smoothing component. Overall, the results are in line with our expectations: as QL and BM25 conditionally satisfy all axioms according to their analytical analyses [13,14] they should satisfy a large percentage of our extended and relaxed axiomatic instances as well. However, these numbers do not reflect the (un)conditional fulfillment of BM25 and QL per original axiom on a one-to-one basis, for which one possible explanation is our relaxation of the document length difference δ.

When we consider the axiomatic scores of our evaluated neural models we observe a clear gap: while for $\overline{\texttt{TFC1}}$ (i.e., documents with more query terms should have higher retrieval scores) between 69-85% instances are satisfied, for $\overline{\texttt{TFC2}}$ (i.e., the increase in retrieval score becomes smaller as the absolute term count increases) and $\overline{\texttt{M-TDC}}$ (i.e., documents with more occurrences of rare query terms are favoured) this drops to at most 76%. When considering the $\overline{\texttt{LNC2}}$ axiom, we find that only aNMM is able to learn the underlying pattern to some degree (38% of satisfied instances) without observing instances of duplicated documents in training ($\overline{\texttt{LNC2}}^{Test}$); the remaining neural models correctly rank between 0 and 19% of instances. Once we include the diagnostic dataset instances in the training regime ($\overline{\texttt{LNC2}}^{All}$) all models have learned to some degree that duplicated document content should not be penalized, but still, none of the models is able to satisfy even half of the diagnostic instances. Finally, we note that aNMM achieves a higher retrieval effectiveness than QL, while QL outperforms aNMM across all four diagnostic datasets. This is an indication that fulfillment of those four axioms alone is not a perfect indicator of retrieval effectiveness—after all, more than twenty have been proposed in the literature. We leave the evaluation of additional axioms to future work. The correlation between retrieval effectiveness in MAP and the average axiomatic score across all axioms is 0.44 ($N = 6$ retrieval models); this is a positive trend, but not a significant one due to the overall low number of models compared.

What we have gained are insights into the type of patterns our neural models have (not) learned and can use those insights to "fix" the models, just like the traditional IR models were fixed based on their axiomatic analyses. As an example, we may want to train Duet on additional triplets, $\mathbf{q}, \mathbf{d_i}, \mathbf{d_j}$, for which $S(\mathbf{q}, \mathbf{d_i}) > S(\mathbf{q}, \mathbf{d_j})$ according to $\overline{\texttt{TFC1}}$, as Duet currently performs worst on this axiom across the evaluated models.

Table 2: Overview of models' retrieval effectiveness and fraction of fulfilled axiom instances. $^{1/2/3/4}$ denote statistically significant improvements (Wilcoxon signed rank test with $p < 0.05$) in retrieval effectiveness.

	Retrieval effectiveness			**Performance per axiom**				
	MAP	MRR	P@5	$\overline{\texttt{TFC1}}$	$\overline{\texttt{TFC2}}$	$\overline{\texttt{M-TDC}}$	$\overline{\texttt{LNC2}}^{Test}$	$\overline{\texttt{LNC2}}^{All}$
[1] BM25	$0.52^{3,4}$	$0.60^{3,4}$	0.18^{3}	0.73	**0.98**	**1.00**	**0.80**	**0.80**
[2] QL	$\mathbf{0.54}^{1,3,4}$	$\mathbf{0.62}^{1,3,4}$	$\mathbf{0.19}^{3}$	**0.87**	0.63	**0.94**	0.68	0.68
[3] Duet	0.25	0.29	0.10	0.69	0.56	0.48	0.19	0.47
[4] MatchPyramid	0.44^{3}	0.51^{3}	0.18^{3}	0.79	0.58	0.63	0.00	0.19
[5] DRMM	$0.55^{1,2,3,4}$	$0.64^{1,2,3,4}$	$0.20^{1,2,3,4}$	0.84	**0.60**	**0.76**	0.05	0.12
[6] aNMM	$\mathbf{0.57}^{1,2,3,4}$	$\mathbf{0.66}^{1,2,3,4}$	$\mathbf{0.21}^{1,2,3,4}$	**0.85**	0.56	0.69	**0.38**	**0.47**

5 Conclusions

In this paper, we have proposed a novel approach to empirically analyze retrieval models that is rooted in the axiomatic approach to IR. Today's neural models, with potentially millions of parameters are too complex for any kind of analytical evaluation; instead, we take inspirations from the NLP and computer vision communities and propose the use of model-agnostic *diagnostic datasets* in order to determine what kind of search heuristics neural models are able to learn. We have shown for four specific axioms how to extend and relax them, in order to make them match realistic datasets. We have applied our diagnostic dataset creation pipeline to the `WikiPassageQA` corpus and evaluated two traditional baselines and four neural models. As a model's axiomatic performance does not require a labelled dataset (i.e., no relevance judgments are required), we can apply our pipeline to almost any dataset containing queries and documents.

Our future work will extend this work in several directions: we will (i) investigate the impact of the adopted document length (δ) relaxation; (ii) extend and relax additional axioms to enlarge our set of diagnostic datasets; (iii) empirically evaluate a larger set of neural models and subsequently attempt to "fix" them (through training data augmentation or the adaptation of their loss function); ad (iv) evaluate a wider range of datasets in order to determine the impact of the retrieval task on the models' axiomatic performance.

Overall, we believe that the axiomatic approach to diagnosing neural IR models presented in this work is a step forward to gaining valuable insights into the black boxes that deep models are generally considered to be.

Acknowledgements: This work was funded by NWO projects LACrOSSE (612.001.605) and SearchX (639.022.722) and Deloitte NL.

References

1. Ariannezhad, M., Montazeralghaem, A., Zamani, H., Shakery, A.: Improving retrieval performance for verbose queries via axiomatic analysis of term discrimination heuristic. In: SIGIR '17. pp. 1201–1204 (2017)

2. Chen, H., Han, F.X., Niu, D., Liu, D., Lai, K., Wu, C., Xu, Y.: MIX: Multi-Channel Information Crossing for Text Matching. In: KDD '18. pp. 110–119 (2018)
3. Clinchant, S., Gaussier, E.: Is document frequency important for PRF? In: ICTIR '11. pp. 89–100 (2011)
4. Clinchant, S., Gaussier, E.: A theoretical analysis of pseudo-relevance feedback models. In: ICTIR '13. pp. 6–13 (2013)
5. Cohen, D., O'Connor, B., Croft, W.B.: Understanding the Representational Power of Neural Retrieval Models Using NLP Tasks. In: ICTIR '18. pp. 67–74 (2018)
6. Cohen, D., Yang, L., Croft, W.B.: WikiPassageQA: A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval. In: SIGIR '18. pp. 1165–1168 (2018)
7. Craswell, N., Croft, W.B., de Rijke, M., Guo, J., Mitra, B.: SIGIR 2017 Workshop on Neural Information Retrieval. In: SIGIR '17. pp. 1431–1432 (2017)
8. Craswell, N., Croft, W.B., de Rijke, M., Guo, J., Mitra, B.: Report on the Second SIGIR Workshop on Neural Information Retrieval. SIGIR Forum **51**(3) (2018)
9. De Boom, C., Van Canneyt, S., Demeester, T., Dhoedt, B.: Representation learning for very short texts using weighted word embedding aggregation. Pattern Recognition Letters **80**, 150–156 (2016)
10. Fan, Y., Guo, J., Lan, Y., Xu, J., Zhai, C., Cheng, X.: Modeling diverse relevance patterns in ad-hoc retrieval. In: SIGIR '18. pp. 375–384 (2018)
11. Fan, Y., Pang, L., Hou, J., Guo, J., Lan, Y., Cheng, X.: MatchZoo: A Toolkit for Deep Text Matching. arXiv preprint arXiv:1707.07270 (2017)
12. Fang, H.: A re-examination of query expansion using lexical resources. ACL HLT '08 pp. 139–147 (2008)
13. Fang, H., Tao, T., Zhai, C.: A formal study of information retrieval heuristics. In: SIGIR '04. pp. 49–56 (2004)
14. Fang, H., Tao, T., Zhai, C.: Diagnostic Evaluation of Information Retrieval Models. ACM Trans. Inf. Syst. **29**(2), 7:1–7:42 (2011)
15. Fang, H., Zhai, C.: An exploration of axiomatic approaches to information retrieval. In: SIGIR '05. pp. 480–487 (2005)
16. Fang, H., Zhai, C.: Semantic term matching in axiomatic approaches to information retrieval. In: SIGIR '06. pp. 115–122 (2006)
17. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: CIKM '16. pp. 55–64 (2016)
18. Hagen, M., Völske, M., Göring, S., Stein, B.: Axiomatic result re-ranking. In: CIKM '16. pp. 721–730 (2016)
19. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: NIPS '14. pp. 2042–2050 (2014)
20. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: CIKM '13. pp. 2333–2338 (2013)
21. Hui, K., Yates, A., Berberich, K., de Melo, G.: Co-PACRR: A Context-Aware Neural IR Model for Ad-hoc Retrieval. In: WSDM '18. pp. 279–287 (2018)
22. Jia, R., Liang, P.: Adversarial Examples for Evaluating Reading Comprehension Systems. In: EMNLP '17. pp. 2021–2031 (2017)
23. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR '17. pp. 1988–1997 (2017)
24. Karimzadehgan, M., Zhai, C.: Axiomatic analysis of translation language model for information retrieval. In: ECIR '12. pp. 268–280 (2012)

25. Li, C., Sun, Y., He, B., Wang, L., Hui, K., Yates, A., Sun, L., Xu, J.: NPRF: A Neural Pseudo Relevance Feedback Framework for Ad-hoc Information Retrieval. In: EMNLP '18. pp. 4482–4491 (2018)
26. Lu, Z., Li, H.: A deep architecture for matching short texts. In: NIPS '13. pp. 1367–1375 (2013)
27. Lv, Y., Zhai, C.: Lower-bounding term frequency normalization. In: CIKM '11. pp. 7–16 (2011)
28. Mitra, B., Craswell, N.: An Introduction to Neural Information Retrieval. Foundations and Trends in Information Retrieval **13**(1), 1–126 (2018)
29. Mitra, B., Diaz, F., Craswell, N.: Learning to match using local and distributed representations of text for web search. In: WWW '17. pp. 1291–1299 (2017)
30. Montazeralghaem, A., Zamani, H., Shakery, A.: Axiomatic analysis for improving the log-logistic feedback model. In: SIGIR '16. pp. 765–768 (2016)
31. Na, S.H.: Two-stage document length normalization for information retrieval. TOIS '15 **33**(2), 8:1–8:40 (2015)
32. Nie, Y., Li, Y., Nie, J.Y.: Empirical Study of Multi-level Convolution Models for IR Based on Representations and Interactions. In: ICTIR '18. pp. 59–66 (2018)
33. Onal, K.D., Zhang, Y., Altingovde, I.S., Rahman, M.M., Karagoz, P., Braylan, A., Dang, B., Chang, H.L., Kim, H., McNamara, Q., et al.: Neural information retrieval: At the end of the early years. Information Retrieval Journal **21**(2-3), 111–182 (2018)
34. Pang, L., Lan, Y., Guo, J., Xu, J., Cheng, X.: A Study of MatchPyramid Models on Ad-hoc Retrieval. arXiv preprint arXiv:1606.04648 (2016)
35. Pang, L., Lan, Y., Guo, J., Xu, J., Cheng, X.: A Deep Investigation of Deep IR Models. arXiv preprint arXiv:1707.07700 (2017)
36. Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S., Cheng, X.: Text Matching as Image Recognition. In: AAAI '16. pp. 2793–2799 (2016)
37. Pang, L., Lan, Y., Guo, J., Xu, J., Xu, J., Cheng, X.: DeepRank: A New Deep Architecture for Relevance Ranking in Information Retrieval. In: CIKM '17. pp. 257–266 (2017)
38. Rao, J., Yang, W., Zhang, Y., Ture, F., Lin, J.: Multi-Perspective Relevance Matching with Hierarchical ConvNets for Social Media Search. arXiv preprint arXiv:1805.08159 (2018)
39. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: Learning semantic representations using convolutional neural networks for web search. In: WWW '14. pp. 373–374 (2014)
40. Shi, S., Wen, J.R., Yu, Q., Song, R., Ma, W.Y.: Gravitation-based model for information retrieval. In: SIGIR '05. pp. 488–495 (2005)
41. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: International Conference on Intelligence Analysis (2004)
42. Tao, T., Zhai, C.: An exploration of proximity measures in information retrieval. In: SIGIR '07. pp. 295–302 (2007)
43. Wan, S., Lan, Y., Guo, J., Xu, J., Pang, L., Cheng, X.: A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations. In: AAAI '16. pp. 2835–2841 (2016)
44. Weston, J., Bordes, A., Chopra, S., Rush, A.M., van Merriënboer, B., Joulin, A., Mikolov, T.: Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. arXiv preprint arXiv:1502.05698 (2015)
45. Yang, L., Ai, Q., Guo, J., Croft, W.B.: aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model. In: CIKM '16. pp. 287–296 (2016)

46. Yang, Y., Yih, W.t., Meek, C.: WikiQA: A challenge dataset for open-domain question answering. In: EMNLP '15. pp. 2013–2018 (2015)
47. Yang, Z., Lan, Q., Guo, J., Fan, Y., Zhu, X., Lan, Y., Wang, Y., Cheng, X.: A Deep Top-K Relevance Matching Model for Ad-hoc Retrieval. In: CCIR '18. pp. 16–27 (2018)
48. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: SIGIR '01. pp. 334–342 (2001)
49. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: NIPS '15. pp. 649–657 (2015)
50. Zoph, B., Le, Q.V.: Neural Architecture Search with Reinforcement Learning. arXiv preprint arXiv:1611.01578 (2016)