Abstract

Although recent years have been marked by incredible advances in the whole development process of NLP systems, there are still blind spots in characterizing what is still hampering real-world adoption of models in knowledge-intensive settings. In this paper, we illustrate through a real-world zero-shot text search case for information seeking in scientific papers, the masked phenomena that the current process of measuring performance might not reflect, even when benchmarks are, in appearance, faithfully representative of the task at hand. In addition to experimenting with TREC-COVID and NFCorpus, we provide an industrial, expert-carried/annotated, case of studying vitamin B’s impact on health. We thus discuss the misalignment between solely focusing on single-metric performance as a criterion for model choice and relevancy as a subjective measure for meeting a user’s need.

1 Introduction

Scientific publications are one of the primary means by which researchers disseminate their findings and discoveries to the community, but the amount of information to go through can easily become daunting and challenging. Exploratory search, i.e., the process of conducting broad and open-ended searches to gain a better understanding of a research topic, is the type of search task that scientists typically spend the most time on. Unfortunately, specialized search engines designed for scientists only partially support this type of task, leaving researchers with limited options for efficiently accessing and extracting relevant information from the vast amount of available literature. We are however currently standing at an age of redefining the way we seek information as we make great advances in the whole development cycle of NLP technologies aiming to solve knowledge-intensive tasks. To this end, benchmarks constitute the backbone of this process and fundamentally influence the way we measure progress and identify where future research efforts should be focused. These datasets, almost solely, put the emphasis on performance-driven comparison and create an impression of reliable estimates of progress at a task scale, whereas, in reality, they might not be informative about the way the models would solve human problems or help solve them.

In this paper, we illustrate an example of transitioning from two biomedical IR benchmarks, i.e., NFCorpus and TREC-COVID, to a practical case of seeking information about a specific topic in scientific publications: vitamin B’s impact on human health. In this context, we provide an expert-annotated collection of relevance judgements on 1811 publications related to vitamin B and health. Our goal is to assess how models’ comparison on benchmarks is meaningful to solving/assisting the expert seeking such information. We thus show through a zero-shot setting that narrowing the comparison down to a single metric might not be relevant to users’ needs, even when a real-world case presents similarities with widely-used benchmarks.

Our contributions can be summarized as follows: 1. we provide a real-world case of information-seeking in scientific publications that does not drift away from prominent benchmarks’ characteristics, 2. we test in a zero-shot setting a few SOTA models reflecting the current paradigm in NLP and IR and we give an interpretation of their behavior in our case compared to the benchmarks, and 3. we discuss based on our observations the masked phenomena that the current process of evaluation might not reflect.

2 Background

2.1 Information-seeking and relevance

Information-seeking strategies described by Belkin et al. (1995) represent how a searcher might use
different methods and resources and have different aims. The broad information-seeking behavior of scientists is usually regarded as an exploratory search problem (Meho and Tibbo, 2003; Athukorala et al., 2013; Nedumov and Kuznetsov, 2019). When searching for information in scientific publications, experts have specific information needs, and they seek information that is relevant to those needs. Effective search systems aim to retrieve highly relevant information and present it to the user in a way that is easy to understand and use. To this end, relevance refers to the degree to which a piece of information satisfies the information need of the seeker. There is however a certain degree of variability in the perception of relevance for a given task (Soufan et al., 2022). Previous work argues that there is a strong relationship between the task singularity, the task carrier, and the type of expected relevance, leading to significant variability of performance levels across tasks (Zhang, 2014; Tamine et al., 2015; Hoeber and Storie, 2022). Relevance can therefore be considered as a subjective measure that depends on both the information-seeker as well as the environment they seek in.

2.2 Semantic Search on Scientific Publications

In the first steps of information seeking in scientific papers, the user may not have a clear understanding or ability to precisely articulate their information need (Vakkari, 2005) thus requiring a search system to understand the meaning behind the query and the content of the documents, rather than just matching the query terms with the terms in the papers. To this end, semantic search on scientific papers refers to the ability of a system to understand the meaning of the query and the content of the scientific papers and match them based on their semantic similarity.

In practice, given a collection of scientific papers $C$, the goal is to rank the most relevant subset of candidate papers $R \subseteq C$ by relatedness to $q$, i.e., $R = \{ p \in C \mid \text{relevance}_q(p) \}$. A common approach to measuring the similarity between the query $q$ embeddings and candidate paper $p$ embeddings. The papers with the highest similarity scores are considered the most similar to the query and are thus returned at the top of the search results.

Among the recent research directions, there has been a focus on learning representations of scientific documents that can be used as rich input features for downstream tasks, thus alleviating the need for further fine-tuning (Cohan et al., 2020; Parisot and Zavrel, 2022; Singh et al., 2022). Zero-shot robustness directions continued to show promising results as well, with state-of-the-art being dominated by models optimized to resist to natural dataset shifts (Yu et al., 2022).

2.3 Evaluation Paradigm

Benchmarks are designed to replicate tasks and are useful for providing a standard method of comparison, reproducibility, and a concise way of tracking progress. For search on scientific papers, a widely adopted paradigm is to provide relevance annotations and evaluate model performance with top-k metrics, notably the Normalised Cumulative Discount Gain @k (Wang et al., 2013) which provides a good balance suitable for tasks involving binary and graded relevance judgements. Nonetheless, there are some concerns with this evaluation methodology.

It has long been argued that information seeking/retrieval is or should be considered as an interactive process with human involvement (Cool and Belkin, 2002; Järvelin, 2011; Shah and Bender, 2022) where a user is more likely to navigate through different information-seeking strategies during a search session (Hoeber et al., 2019). Current benchmarks are however non-interactive, whereas the information-seeking process is or should be considered an interactive process with human involvement. To this end, models that are deployed for interactive purposes should be evaluated as such (Lee et al., 2022).

Further, top-k metrics assume that lower ranks are not of interest, with benchmarks usually evaluating on k=10. This contributed to favoring speed and convenience, but in such knowledge-intensive settings like searching about a particular topic in scientific papers, the priority is to fill in the gaps of knowledge of the information-seeker (Hasan Awadallah et al., 2014). Such small values of $k$ present a very strong assumption on the quantity of information an expert requires to study their subject. We argue in the rest of the paper that such a method may not be ideal for evaluating systems that involve users in expert search situations, as it may not fully account for factors such as the user’s
interests and expertise when assessing relevance.

3 Experimental Setup

In this section, we provide a description of our experimental setting of transitioning from information-seeking benchmarks on scientific papers to an industrial exploratory search about vitamin B’s impact on health.

3.1 Datasets

TREC-COVID (Voorhees et al., 2021) is a test collection leveraging the TREC framework that aimed to address the search needs of clinicians and biomedical researchers during the COVID-19 pandemic, thus laying the groundwork for improved search systems in future public health emergencies. The document set is the one provided by CORD-19 (Wang et al., 2020), which consists of new and historical publications on coronaviruses. TREC-COVID aims to rank papers in response to textual search queries and contains 50 queries with 69,318 candidate papers cumulatively judged by relevance.

NFCorpus (Boteva et al., 2016) is a Medical Information Retrieval data set on nutrition facts where queries are harvested from NutritionFacts.org site and candidate documents are medical research papers mostly from PubMed. We use the dataset as it is contained in the BEIR benchmark (Thakur et al., 2021): 323 queries and 3633 candidate documents.

Practical case: Vitamin B’s impact on health

We present a practical case study where an expert in immunology seeks to study the effects of vitamin B on human health. A corpus of candidate papers was retrieved from PubMed with the following query: (“vitamin B”[Title/Abstract]) AND (health[Title/Abstract] OR growth[Title/Abstract]), which resulted in 1811 papers, out of which the expert identified 598 relevant documents (33%). Relevance judgement was carried out in two steps: 1. Search on title relevance: if a title is obviously out of scope, the expert does not investigate the abstract. Similarly, if the title is evidently in scope, the abstract is not judged. 2. Search on abstract relevance: the expert reads in detail and identifies the type of study that was carried out. On the models’ side, the query used for ranking is: How do vitamins B impact health?

Our vitamin B case has some similarities in nature with both NFCorpus and TREC-COVID (although not identical). While identifying and analyzing discrepancies between benchmarks and use-case results can provide valuable insights for improving the performance of models in practical real-world use, it can be difficult to know for certain whether a benchmark is representative of a real-world task, as this requires a careful investigation of the data, input format, expert input, and evaluation metrics.

3.2 Models & Frameworks

Transformer-based models have gained widespread popularity as retrieval models, due to their capability of acquiring semantic representations. We use BM25 as a generalizable baseline (Thakur et al., 2021) and test two sets of neural models that we port to sentence-transformers (Reimers and Gurevych, 2019) format known for its efficiency in semantic search tasks (Muennighoff, 2022):

1. LMs pre-trained for scientific text similarity: SPECTER (Cohan et al., 2020), SciNCL (Osendorff et al., 2022), and ASPIRE (Mysore et al., 2022). All three have been trained with the intuition that learned scientific document representations can be substantially improved through contrastive learning objectives.

2. Robust models in zero-shot settings, COCODR (Yu et al., 2022) and monoT5 (Nogueira et al., 2020), both transferred from MS-MARCO.

Finally, we use Haystack as a framework and ElasticSearch to index embeddings along the papers. We did not alter the original trainings of models.

4 Results & Discussion

We report in Table 1 the average nDCG@10 of the different models on both NFCorpus and TREC-COVID, as well as our use case. We experiment with three strategies of searching: based on title relevance, on abstract relevance, and titles and abstracts appended.

NFCorpus BM25 is leading on the three strategies, followed by scientific LMs mostly dominating the general robust models. The low scores (compared to the other datasets) can be partially explained with the fact that the percentage of relevant articles is smaller for most queries. All models, with the exception of BM25 and

---

3 Retrieved in December 2022. All papers are in English.
4 https://microsoft.github.io/msmarco/
5 https://github.com/deepset-ai/haystack
6 Separated with [SEP] token.
Table 1: Average NDCG@10, * denotes the best score while _ denotes the worst performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Title</th>
<th>Abstract</th>
<th>T+Abs</th>
<th>Title</th>
<th>Abstract</th>
<th>T+Abs</th>
<th>Title</th>
<th>Abstract</th>
<th>T+Abs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>0.335</td>
<td>0.375</td>
<td>0.380</td>
<td>0.579</td>
<td>0.646</td>
<td>0.659</td>
<td>0.496</td>
<td>0.06</td>
<td>0.066</td>
</tr>
<tr>
<td>SPECTER</td>
<td>0.155</td>
<td>0.156</td>
<td>0.161</td>
<td>0.654</td>
<td>0.631</td>
<td>0.66</td>
<td>0.621</td>
<td>0.402</td>
<td>0.77</td>
</tr>
<tr>
<td>SciNCL</td>
<td>0.207</td>
<td>0.182</td>
<td>0.195</td>
<td>0.68</td>
<td>0.635</td>
<td>0.657</td>
<td>0.534</td>
<td>0.637</td>
<td>0.70</td>
</tr>
<tr>
<td>ASPIRE</td>
<td>0.216</td>
<td>0.188</td>
<td>0.193</td>
<td>0.688</td>
<td>0.672</td>
<td>0.685</td>
<td>0.536</td>
<td>0.445</td>
<td>0.546</td>
</tr>
<tr>
<td>COCO-DR</td>
<td>0.209</td>
<td>0.127</td>
<td>0.139</td>
<td>0.72</td>
<td>0.654</td>
<td>0.714</td>
<td>1.0</td>
<td>0.748</td>
<td>0.848</td>
</tr>
<tr>
<td>monoT5</td>
<td>0.114</td>
<td>0.044</td>
<td>0.046</td>
<td>0.468</td>
<td>0.512</td>
<td>0.513</td>
<td>0.538</td>
<td>0.863</td>
<td>0.87</td>
</tr>
</tbody>
</table>

SPECTER, perform better on titles rather than assessing abstracts’ relevance.

**TREC-COVID** On titles and titles+abstracts, COCO-DR is the best-performing model, whereas ASPIRE slightly outperforms it on abstracts. Models’ performances are quite consistent on this benchmark, with scientific LMs having close scores and mostly best performing on titles. TREC-COVID’s queries are more detailed than the other two datasets. This might explain the coherence of results between models; there is more relevant information to judge on.

**Vitamin B & Health** Models’ performance in our case seems to be divergent from what can be observed on the other two benchmarks. BM25’s performance entirely drops with abstracts, which might be caused by the nomenclature of vitamin B (Appendix A) present in titles and abstracts. On the other hand, monoT5 outperforms all other models on strategies that include abstract relevance, whereas COCO-DR achieves perfect nDCG@10 on titles.

Overall, our results show that models perform differently on datasets and suggest that there is an inconsistency in performance and difficulty in identifying the best model for seeking biomedical information in publications: if starting from NFCorpus, one would suggest using BM25 as a decent model for the vitamin B case, whereas if comparing on TREC-COVID, one would prefer COCO-DR and entirely leave out monoT5. In reality, the perfect nDCG@10 on titles of COCO-DR might suggest the best fit, but the model is not actually placing all the relevant documents at the smallest ranks: Figure 1 illustrates this and shows that the nDCG@10 metric is not reflecting how "early" relevant documents are suggested to the user (the tendency is the same on TREC-COVID (Appendix B) for SciNCL, COCO-DR, and monoT5). The differences of scores in Table 1 suggest a big gap in the performance of models, however, if we consider the entire set of relevant papers in the vitamin B case, SciNCL (ndcg@10=0.534) is cumulatively suggesting the relevant elements "faster" than COCO-DR (ndcg@10=1.0), making it a better assistance to the expert seeking information.

## 5 Discussion

We further discuss in this section the misalignment between performance measures from the perspective of an expert seeking information in papers.

**Expert search strategy preferences** Our expert expressed that they sort on titles for more speed, but abstract relevance remains the reference. The reason for this is that titles usually provide information on the study domain as a whole, and can be used to classify into big categories. The abstracts however are used when the title does not allow for immediate classification, since they contain the main question of the paper, the methodology, and the main results. Intuitively, models would find more relevance "hints" in abstracts, and thus have greater performance on search strategies that include them.

This was rarely the case for all datasets, suggesting that many models might be better at matching shorter contexts (titles being closer to query length compared to abstracts).

**The success of information-seeking is a process** As we previously mentioned, tasks that are complex, such as learning about a new topic, often require multiple search queries and may extend over multiple sessions. It has to be noted that our ex-
pert encountered 18 different themes out-of-scope (Appendix A) when annotating the entire collection of papers. These themes are discovered during the exploration process, emphasizing the fact that information-seeking is an interactive process and that the reported metric (designed for speed and convenience of ranking systems) is neither informative about the presence of such themes nor about the corresponding response of the different models. As we mentioned in Section 2.1, relevance is a subjective measure. Our expert investigated the ranked lists returned by different models on the vitamin B use case and categorized the first 100 irrelevant documents for each. We observed that the models’ sensitivity to different topics is not the same when measuring similarity. For instance, on titles, COCO-DR (best performance) struggled most with practice recommendations, while SciNCL misjudged the prevalence of B vitamin deficiencies the most. Further, this was also the case on abstracts (Appendix A) as monoT5 struggled most with the vitamin content of food/diet, while BM25 suggested irrelevant studies the most. We illustrate these differences in Figure 2: no agreement whatsoever between models about (ir)relevance of topics, which cannot be reflected by the NDCG@k measure. Such a disagreement further complicates the process of identifying the sources of differences, which are important to determine which model may be better suited for specific scenarios, given that such differences might have roots in the training data, model architectures, hyperparameters, or other factors.

6 Conclusion

In this paper, we illustrated the misalignment between single-metric performance and relevance in practical expert information seeking. Through a transition from two biomedical IR benchmarks to a case of an expert seeking information about vitamin B’s impact on human health, we showed that the current process of measuring performance may not fully capture the challenges of the task at hand. Our observations emphasized the misalignment between relying on top-k ranking metrics and the true nature of the information-seeking process’ success. To this end, we provide an extensive description of the use-case creation and relevance judgements to foster future reconciliation between corpus-based evaluations and users’ search experience.
Limitations

We presented in this paper a real-world annotated example of seeking information in scientific publications. Even if the number of instances presented here is of the same order of magnitude as what is present in benchmarks, we presented only one query and its correspondent relevance judgements, provided by one expert, due to resource constraints. As we noted above, building a corpus dedicated to the exploration of a single information need does however correspond to a real industrial use case. Further, we favored the use of sentence-transformers format for all neural models for the sake of efficiency. We did not dive into providing the best-known performing models and did not consider optimizing them in our case, as overfitting to our data might induce errors in conclusions and low confidence in the generalizability of our observations. However, we do not guarantee that other models will not display more robustness to the transition presented in our paper.

Finally, we did not conduct an extensive examination of the characteristics of the benchmarks as well as the real-world case that may be impacting performance such as the diversity of data. We believe that such investigations, in conjunction with the models’ examination, might help better explain the models’ behaviors and areas of weakness.

Acknowledgements

We would like to thank the reviewers for taking the time to provide such thoughtful and detailed feedback on our work. This work has been funded by the ANRT CIFRE convention N°2019/1314 and ERDYN.

References


**A Vitamin B & Health**

We provide in this appendix further details about our real-world case that were partially discussed in the paper.

**Out-of-scope themes** We list hereafter the out-of-scope themes that were encountered during the annotation process:

- Yeast, bacteria and plants
- Measurement methods of vitamin B.
- Prevalence of deficiencies.
- Which foods bring which vitamin B and in which amounts.
- Recommendations of public health policies.
- Genetic polymorphism leading to different uses of B vitamins.
- Cobalt (essential component of vitamin B12).
- Farm animals (chicken, swine, cattle, fish).
- Interaction of B vitamins with other drugs (such as oral contraceptives).
- B vitamin derivatives (such as Pyridoxal 5'-phosphate, a derivative of B6).
- Nutrient intake from different diets (vegan, vegetarian, omnivorous).
- Effect of surgery on the levels of B vitamins.
- Use of B vitamins to improve in vitro process (such as in vitro growth of follicles).
- Physicians targeted, about how to supplement patients.
- Supplementing people and only looking at biomarkers in response.

**Table 2: Nomenclatures for the different types of B vitamins.**

<table>
<thead>
<tr>
<th>Vitamin</th>
<th>Associated name</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>Thiamin</td>
</tr>
<tr>
<td>B2</td>
<td>Riboflavin</td>
</tr>
<tr>
<td>B3</td>
<td>Niacin / Nicotinamide</td>
</tr>
<tr>
<td>B5</td>
<td>Pantothenic acid / Pantothenate</td>
</tr>
<tr>
<td>B6</td>
<td>Pyridoxine</td>
</tr>
<tr>
<td>B7</td>
<td>Biotin</td>
</tr>
<tr>
<td>B9</td>
<td>Folic acid / Folate</td>
</tr>
<tr>
<td>B12</td>
<td>Cobalamin</td>
</tr>
</tbody>
</table>

Nomenclatures The different names encountered in titles and abstracts associated with vitamin B are detailed in table 2.

**Irrelevant categories** The out-of-scope themes are grouped into the categories described in table 3. The count per category discussed in section 5 is shown on figure 4.

**B TREC-COVID**

Figure 3 shows the count of relevant papers in different ranges of \(k\) as ranked by titles’ relevance. In TREC-COVID, ranges of \(k\) differ for different queries, so we illustrate on the minimum range, *i.e.*, 1842 documents.
<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vitamin B requirement</td>
<td>Studies on the requirement of B vitamins in populations, for instance pregnant women or elderly people. Evolution of the requirements depending on health situation or medication, for example contraceptive pill intake.</td>
</tr>
<tr>
<td>Practice recommendations</td>
<td>Papers on how to manage B vitamin intake or deficiencies, public policies of vitamin reinforcement in food, or specific food intake to cover vitamin requirements.</td>
</tr>
<tr>
<td>Vitamin content of food/diet</td>
<td>Papers measuring or estimating the vitamin content of different foods or diets, for instance the vegetarian or vegan diets.</td>
</tr>
<tr>
<td>Effect of another vitamin</td>
<td>Studies on the effect of vitamins that are not B vitamins.</td>
</tr>
<tr>
<td>Effect in plants</td>
<td>Studies on B vitamins requirements, or effects, or supplementation in terrestrial plants, algae, or plankton.</td>
</tr>
<tr>
<td>Irrelevant model</td>
<td>Effects of B vitamins on health of animal models that are not relevant (all animals except preclinical in vivo studies on mice or rats). Studies on farm animals are excluded.</td>
</tr>
<tr>
<td>Prevalence of B vitamin deficiencies/ measurement of B vitamin in populations</td>
<td>Studies measuring B vitamins in populations.</td>
</tr>
<tr>
<td>Methodology for measurement</td>
<td>Papers describing B vitamin direct measurement tools, or extrapolation methods to infer B vitamin concentration from other parameters.</td>
</tr>
<tr>
<td>Methodology for production</td>
<td>Papers describing production tools, such as genetically modified yeast or bacteria, fermentation, chemical synthesis, and other methods.</td>
</tr>
<tr>
<td>Irrelevant study</td>
<td>Irrelevant but not categorizable.</td>
</tr>
</tbody>
</table>

Table 3: Categories of wrong predictions encountered when analyzing the first 100 irrelevant papers
The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.
C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
   Not applicable. Left blank.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
   Not applicable. Left blank.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, SpaCy, ROUGE, etc.)?
   Section 3.2

D  ✓ Did you use human annotators (e.g., crowdworkers) or research with human participants?
   Section 3.1

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
   Not applicable. Left blank.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants’ demographic (e.g., country of residence)?
   Not applicable. Annotators are among the authors of the paper and the annotation was carried as part of the research work itself.

D3. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
   Not applicable. Left blank.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
   Not applicable. Left blank.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
   Not applicable. Left blank.