Multi-objective Representation Learning for Scientific Document Retrieval

Mathias Parisot
Zeta Alpha
parisot@zeta-alpha.com

Jakub Zavrel
Zeta Alpha
zavrel@zeta-alpha.com

Abstract

Existing dense retrieval models for scientific documents have been optimized for either retrieval by short queries, or for document similarity, but usually not for both. In this paper we explore the space of combining multiple objectives to achieve a single representation model that presents a good balance between both modes of dense retrieval, combining the relevance judgements from MS MARCO with the citation similarity of SPECTER, and the self-supervised objective of independent cropping. We also consider the addition of training data from document co-citation in a sentence context and domain-specific synthetic data. We show that combining multiple objectives yields models that generalize well across different benchmark tasks, improving up to 73% over models trained on a single objective.

1 Introduction

With the explosive growth of the volume of scientific publications, researchers increasingly rely on sophisticated discovery and recommendation tools to find relevant literature and related work (Ammar et al., 2018; Fadaee et al., 2020). In particular, the development of neural information retrieval (Lin et al., 2021) has led to a quest for dense document representations that capture the semantics of documents better than the previous generation of keyword-based retrieval methods. Such representations are typically achieved by specializing pre-trained large language models for the retrieval task. In this paper, we focus on the case of the bi-encoder (Humeau et al., 2019), where at indexing time documents are embedded as dense vector representations and stored in a fast approximate nearest neighbor system, and at retrieval time queries are encoded using the same model and similarity search is performed in the vector space.

The data set that has powered a lot of advances in this area is MS MARCO (Bajaj et al., 2016), which consists of user queries combined with human relevance judgements for documents and passages. Models trained on this data set are the current state of the art for retrieval based on queries, though the discussion is still ongoing about their effectiveness in terms of out-of-domain generalization (Thakur et al., 2021). As Table 1 shows, models based on this data set tend to perform less well on benchmarks that test document-to-document retrieval. In (Cohan et al., 2020), a scientific document retrieval model called SPECTER was introduced that is specifically optimized for document-to-document similarity, based on exploiting the signal in the citation graph between documents. Model trained for document representations tend to perform less well than MS MARCO based models on query-to-document retrieval tasks such as those presented in the BEIR data set (Table 1).

Additional evidence suggests that self-

	BEIR subset	SciDocs	ICLR 2022
Single objective			
MS MARCO	0.270	68.72	0.260
SPECTER	0.207	79.75	0.407
Multi-objective			
ICrop+context2doc	0.285	78.29	0.450
Δ*MS MARCO*	+5.6%	+13.9%	+73.1%
Δ*SPECTER*	+37.7%	-1.8%	+10.6%
AllObj-Alt	0.278	79.44	0.424
Δ*MS MARCO*	+3.0%	+15.6%	+63.1%
Δ*SPECTER*	+34.3%	-0.4%	+4.2%

Table 1: Single objective compared to multi-objective training. Metrics are ndcg@10 for BEIR and ICLR2022 (doc2doc dataset), and average over all tasks for SciDocs. Δ +/- percentages represent the relative improvement compared to single objective models for the given benchmark. ICrop+context2doc is a model trained on independent cropping and finetuned on unarXiv context2doc. AllObj-Alt is trained alternating batches of independent cropping, SPECTER, and MS MARCO .

supervised tasks, such as the Inverse Cloze Task (Lee et al., 2019) or Independent Cropping (Izacard et al., 2021) can make document representations more robust and improve retrieval relevance (Chang et al., 2020; Izacard et al., 2021). Also, in-domain synthetic data has been explored to enhance retrieval effectiveness in new domains (Bonifacio et al., 2022).

So far, however, no systematic study were performed on the combination of multiple objectives for scientific document representation learning. In this paper, we explore several methods to combine different data sets and training objectives and study the effectiveness of these training strategies on a number of scientific document retrieval benchmarks. In addition to this, we introduce and make available a new data set that emphasises document-to-document similarity. By doing this, we try to answer the following research question:

How can we best combine multiple data sets and task objectives to train a scientific document retriever that can be queried both using short queries and documents?

Although in practice a retrieval system could use multiple different document embeddings for multiple tasks, a single multi-purpose document representation offers great advantages in terms of storage space, computational resources, and operational efficiency.

The main contributions of this paper are the following:

1. We train a dense retriever that performs well on both query2doc and doc2doc retrieval;
2. We introduce a new way to use citation context to generate semi-supervision signal for scientific documents;
3. We release a scientific document to document data set with 1844 human annotations among which 441 are positive relevance judgements; and
4. We publish the code and data sets used for our experiments at `https://github.com/zetaalphavector/multi-obj-repr-learning`

2 Related work

Document retrieval. Multiple recent papers focus on learning representations for scientific documents and dense neural document retrieval (Tan et al., 2022; Zhang et al., 2022; Ostendorff et al., 2022a). (Cohan et al., 2020) presents how to use weak supervision from the scientific citation graph to train a dense retrieval model (SPECTER) and introduces SciDocs, a benchmark to evaluate document representations. (Ostendorff et al., 2022b) improves on SPECTER by using a graph embedding model to sample positive and negative documents and create better training triplets. (Abolghasemi et al., 2022) combines a ranking and representation loss to train a query by document retriever. (Althammer et al., 2022) proposes a method to disregard the input length restriction of transformer-based models by using a paragraph aggregation retrieval model. In our own work, we build on (Cohan et al., 2020) and use the same framework but explore how adding multiple training objectives can improve the performance of a document retriever.

citation context. Earlier work by (Colavizza et al., 2017) already shows that co-citation of documents, especially at the sentence level, is a strong signal for semantic relatedness of documents. (Mysore et al., 2022) explores co-citation context supervision for document representation learning, and applies it to aspect matching. We add this signal to our mix of potentially useful constraints in a multi-objective learning setting, and focus on document representation and training models which can be used for both query to document and document to document retrieval. Moreover, we introduce a new co-citation supervision by using the citing sentence context as a query for the documents cited.

3 Method

3.1 Bi-encoder and losses

We focus on dense retrieval using a bi-encoder architecture (Humeau et al., 2019) with a shared encoder E for the query q (here q can be a short query or a document query) and document d. The model E encodes the query and document into representations $E_q = E(q)$ and $E_d = E(d) \in R^n$ respectively. Note that, in our case, the encoder E is a transformer-based model (Vaswani et al., 2017) which means that its output is a sequence of token representations. We aggregate this sequence into a single representation using mean pooling (we also experimented with using the $[CLS]$ token representation without success over mean pooling). The relevance between the query and document is expressed using a distance metric, cosine similarity in our case, between the two representations: $s(q,d) = dist(E_q, E_d)$.

To train the model, we use data sets of triples of the form: (q, d^+, d^-) where d^+ and d^- are documents that are respectively relevant and not relevant for the query q. When possible, we concatenate the title and abstract of a document as follow: $E_d = E(d_{title}[SEP]d_{abstract})$

We experiment with two losses: Multiple Negative Ranking Loss (MNRL) and Triplet Loss (TL). With MNRL, the single negative document d^- is enriched into a set of negative documents D^- composed of the positive and negative documents from the other triples in the training batch.

$$MNRL(q, D) = -\log \frac{\exp\left(s(q, d^+)\right)/\tau}{\sum\limits_{d \in D} \exp\left(s(q, d)\right)/\tau} \tag{1}$$

Where $D = D^- \cup \{d^+\}$ is the set of all positive and negative documents of all the queries in the batch. With TL, each query uses exactly one positive and one negative document.

$$TL(q, d^+, d^-) = \max\left\{d(q, d^+) - d(q, d^-) + \epsilon, 0\right\} \tag{2}$$

Where $d(q,d) = ||E_q - E_d||_2$ is the L2 norm between the representations of the query and document. While several works (Cohan et al., 2020; Ostendorff et al., 2022b) use TL as their loss function,in most of our experiments, MNRL performed better across different data sets and domains. All the presented results in this paper use MNRL.

3.2 Multi-objective training (multiple types of supervision and domain)

Multi-task learning (Caruana, 1993) has shown good results to improve the generalization of language models. Tasks can be machine translations, next-word predictions, information retrieval, and others. Following those advances, we focus here on a single task (information retrieval) but are interested in combining several types of supervision objectives, data sets, and data domains with the goal to increase the amount of useful training signals and to train a more general-purpose dense retriever.

We experiment with multiple types of supervision: fully supervised data, weakly supervised, and self-supervised training. We also explore two domains: online question answering and scientific document representation for finding related documents. The goal of those experiments is to study whether combining multiple objectives can improve performance across the board. Here, an objective refers to a type of supervision combined with a domain.

We start by considering the following three objectives:

- **MS MARCO data** as fully supervised out-of-domain (question answering) data.
- **Scientific citation graph data** as weakly supervised data, automatically extracted from the scientific literature. There are several ways to use the citation graph as supervision signals. A common approach to derive relevance information is to use cited documents as positive examples (Cohan et al., 2020). We also explore other ways such as using co-cited documents within a given citation context and using the context as the query for the documents that it cites.
- **Unsupervised data** generated via independent-cropping (Izacard et al., 2021) on a scientific corpus. Given a document, independent-cropping samples two independent spans of tokens (which can overlap) forming the query and the positive document. The negative document is sampled similarly from another document in the corpus. The original authors suggest that the overlap between query and documents encourages the model to learn lexical matching between query and document. In our implementation, both the query and document spans contain at least ten tokens.

3.3 Combining objectives

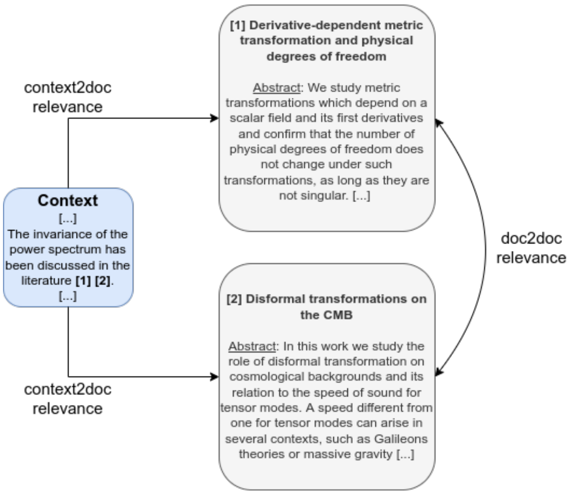
This subsection describes three ways to combine objectives.

In-batch mixing. One way to combine objectives is in-batch objective mixing. Here, data from different objectives are randomly mixed within the same batch. Each of the N objectives is assigned a weight w_i ($\sum_{n=1}^{N} w_i = 1$). A batch of B instances is composed of $w_i \times B$ instances on average of a given objective i. We experiment with multiple weighting configurations. When using MNRL, because the negative documents are shared within the batch, in-batch mixing negatives are more diverse compared to the two other ways to combine objectives.

Alternate batch mixing. Another way to combine objectives is to change the objective for each training iteration. Overall, the training data is equally distributed across the objectives, but there is no mix within a given batch. As opposed to in-batch mixing, the set of negative documents comes from the same objective.

Finetuning. The last way we explore is finetuning. For a given objective A, we train a model on A until convergence and then finetune it on a target objective B. The training on the second objective is shorter and commonly uses a lower learning rate.

4 Experimental Setups



[...]
The invariance of the power spectrum has been discussed in the literature **[1] [2]**.
[...]

[1] Derivative-dependent metric transformation and physical degrees of freedom

Abstract: We study metric transformations which depend on a scalar field and its first derivatives and confirm that the number of physical degrees of freedom does not change under such transformations, as long as they are not singular. [...]

[2] Disformal transformations on the CMB

Abstract: In this work we study the role of disformal transformation on cosmological backgrounds and its relation to the speed of sound for tensor modes. A speed different from one for tensor modes can arise in several contexts, such as Galileons theories or massive gravity [...]

Figure 1: Descriptions of two ways to extract relevant pairs of text for citation contexts. doc2doc uses citation contexts to associated co-cited documents. context2doc uses the context as the query for a cited document.

This section describes details about our experimental setups. We discuss the training and evaluation data sets as well as our choice of hyperparameters.

4.1 Training data

MS MARCO (Bajaj et al., 2016) is a large-scale information retrieval data set created from Bing's search query logs. Sentence-BERT (Reimers and Gurevych, 2019) provides a data set of hard negatives[1] mined from dense models for this data set. We create triplets (q, p, n) where q is the query, p is the annotated positive document, and n is a negative document sampled from the data set. For our experiments, we mined 5 negative documents per system, all with a cross-encoder score of 3.0 or less.

SPECTER, a data set extracted from the Semantic Scholar corpus (Ammar et al., 2018), a data set of scientific papers. We train our models with the subset of the corpus used by Cohan et al.. The data set is composed of triplets (q, p, n) where q is the query paper, p is a paper cited by p, and n is a paper not cited by q but cited by a paper cited by the query paper q. The data set contains 684,100 training triplets and 145,375 validation triplets.

unarXiv (Saier and Färber, 2020) is a large scholarly data set with annotated in-text citations. From it, we extract all one-sentence contexts containing at least two arXiv papers and only select the contexts with citing papers that are posted on arXiv (with an associated arXiv identifier). Our final data set contains a collection 343,578 one-sentence context citing a total of 300,736 arXiv papers across multiple scientific fields (see Figure 2). The contexts and documents contain respectively 29.8 and 152.5 words on average. We use unarXiv for 2 tasks (see Figure 1). First, we use the co-cited documents as positive examples in a document-to-document retrieval setup (we refer to this objective as *doc2doc*). Then, we use the context as the query for any of the documents it cites. We refer to this objective as *context2doc* short for context to document. Our version of the dataset is available here [2].

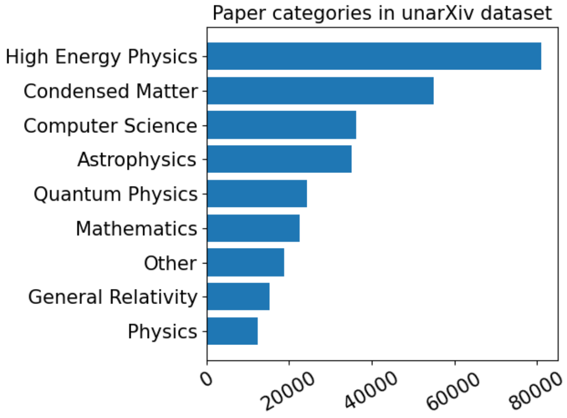


Figure 2: Counts of ArXiv categories for the documents in unarXiv collection. Categories with less than 10,000 documents are grouped into "Other".

InPars (Bonifacio et al., 2022) is a recent method to generate synthetic training data sets for information retrieval tasks. The idea is to use large language models, such as GPT-3 (Brown et al., 2020), to generate queries that are relevant to a

[1] https://huggingface.co/datasets/sentence-transformers/msmarco-hard-negatives

[2] [github link]

given document. Bonifacio et al. generated synthetic data for all the data sets present in BEIR. We combine all the synthetic data sets [3] into one and use it as training data. We only experiment with this data set for finetuning models pre-trained on other objectives.

4.2 Evaluation

BEIR (Thakur et al., 2021) is a benchmark containing 15 information retrieval tasks. We select 5 openly available long document data sets from the benchmark: SciDocs (Cohan et al., 2020); NFCorpus (Boteva et al., 2016) a medical information retrieval data set of 3,244 queries and 9,964 documents; SciFact (Wadden et al., 2020) a scientific fact-checking data set of 300 queries and 5,183 documents; TREC-COVID (Voorhees et al., 2020) a pandemic information retrieval data set of 50 queries and 171,332 documents; ArguAna (Wachsmuth et al., 2018) a counter-argument data set of 1,406 queries and 8,670 documents. Except for ArguAna, where queries have 192.98 words on average, all the other selected BEIR data sets are short queries to document retrieval tasks. The selected data set with the longest queries is SciFact with 12.37 words per query on average.

SciDocs (Cohan et al., 2020) is a framework evaluating scientific paper embeddings. It is composed of 4 tasks: document classification, citation prediction, user activity, and recommendation. Note that the SciDocs task presented above in the BEIR benchmark is only a subtask.

ICLR2022 . Furthermore, we introduce a new specialized document to document retrieval data set of artificial intelligence scientific papers. We create our corpus from all the 1094 papers presented at ICLR 2022 [4]. We randomly sample 40 of those papers and use them as our queries. We index the corpus using FAISS (Johnson et al., 2019) library and retrieve a list of 10 documents with cosine similarity using multiple models. We distribute the query-document pairs across 4 in-house annotators and manually annotate the pairs using 3-scale relevance judgements: 0 not-relevant, 1 relevant and 2 very relevant. Removing the duplicate pairs across the ranking lists of different models, the data set contains 1,844 relevance judgements out of which 358 are relevant and 83 are very relevant. The dataset is available here [5].

4.3 Hyper-parameters and training details

We use a pre-trained MiniLM-L6[6] Transformer model as a basis, and train each of our models from this for a maximum of 200,000 steps or until convergence of the validation loss with a patience of 2. Each training batch contains 16 triplets and we accumulate the gradients during 2 steps. When multiple objectives are combined during training, the convergence metric is the average of the validation losses of the training objectives. The optimizer is AdamW (Loshchilov and Hutter, 2017) with a learning rate of 2×10^{-5}, no weight decay and $\epsilon = 10^{-8}$. The learning rate follows a linear schedule without warmup. When finetuning models on a target objective, we train the model on a single epoch of the target data set using a learning rate of 10^{-5}. The rest of the optimizer and scheduler parameters stay the same.

All the experiments were run on a single NVDIA Titan RTX GPU with 24GB GDDR6. The use of the MiniLM-L6 model means that we were able to do fast experiments, but also that the results we report in this paper are not directly comparable to the state-of-the-art achieved using much larger models.

5 Results and Discussion

Single vs. Multi-objective training. We first study the impact of adding supervision signals from multiple sources compared to a single training objective. Table 1 presents the results of *ICrop. + context2doc* and *AllObj-Alt* two multi-objective models compared to the best single-objective models on each of the 3 evaluation metrics. Both multi-objective models manage to outperform the baseline model trained on MS MARCO for all metrics with at least 3% and up to 38% improvement on single metrics. The multi-objective models reached performance on SciDocs close to the baseline model trained on SPECTER while toping its score on BEIR and ICLR2022 . We take those results as empirical evidence that multi-objective training leads to models that generalize better across multiple domains.

Combining objectives. We study the impact of the four combining methods: in-batch mixing,

[3] https://github.com/zetaalphavector/inPars

[4] https://iclr.cc/Conferences/2022

[5] [link to dataset]

[6] https://huggingface.co/nreimers/MiniLM-L6-H384-uncased

Training objectives	Combining	BEIR-subset	SciDocs	ICLR2022
		(ndcg@10)	(avg.)	(ndcg@10)
2 objectives				
(1) MS MARCO , (2) ICrop.	in batch mix	0.269	74.93	0.342
	alternate	0.282	74.51	0.341
	finetune 1 → 2	0.262	74.27	0.343
	finetune 2 → 1	**0.324**	**75.09**	**0.396**
(1) SPECTER, (2) MS MARCO	in batch mix	0.230	78.90	0.396
	alternate	0.258	**79.25**	0.381
	finetune 1 → 2	**0.307**	75.08	0.330
	finetune 2 → 1	0.265	78.91	**0.419**
(1) SPECTER, (2) ICrop.	in batch mix	0.127	78.85	0.413
	alternate	0.239	78.57	0.384
	finetune 1 → 2	0.242	75.78	0.375
	finetune 2 → 1	**0.248**	**79.40**	**0.471**
3 objectives				
MS MARCO , SPECTER, ICrop.	in batch mix	0.244	77.78	0.389
	alternate	**0.278**	**79.44**	**0.424**

Table 2: Comparison of combining objectives methods. BEIR subset is the average ndcg of the 5 tasks, SciDocs avg is the average metric of all the tasks. Results in bold are the best result given a set of objectives, underlined results are the best overall.

alternate, and finetuning in both directions. To do so we combine three objectives: MS MARCO , SPECTER, and independent cropping. The results are presented in Table 2. When using only two data sets, finetuning is a better option than both in-batch mixing and alternating between batches. The results are consistent across the 3 pairs of objectives. There is no clear preference between alternating batches and in-batch mixing for two objectives. The models trained with the latter perform best on SciDocs and ICLR2022 while the models trained with alternate batches perform best on BEIR. Maybe the diverse domains of the BEIR data sets create negatives that are too easy to spot, while the domains of SciDocs and ICLR2022 are relatively similar making the negative harder and forcing the model to learn better representations. When using three objectives, alternating between batches performs well across the three evaluation metrics and seems like the best compromise.

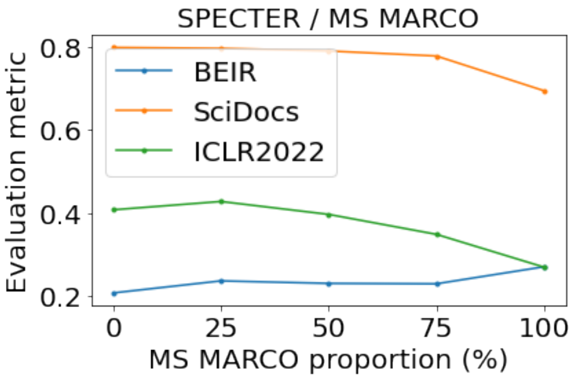


Figure 3: Performance on SciDocs, BEIR, and ICLR2022 when combining SPECTER and MS MARCO objectives with in-batch mixing and varying the proportion of data instances coming from MS MARCO . When MS MARCO proportion is 25% the split is 75%-25%. The evaluation metric for SciDocs is divided by 100 for clarity.

Split proportion. We analyze the effect of the split proportion when combining 2 objectives using in-batch mixing. Figure 3 presents the performances of a model trained using in-batch mixing on MS MARCO and SPECTER objectives. We explore 5 split-proportions going from a model trained on only SPECTER data (0% MS MARCO) to one trained using 100% MS MARCO data. The figure shows that only adding a small proportion of MS MARCO data results in better performance on BEIR while maintaining high performance on SciDocs and ICLR2022 (25% MS MARCO gives the highest score on ICLR2022). As expected, training on a larger proportion of MS MARCO increases the performance on BEIR but the trade-off is not interesting as the document representation and document retrieval performance suffer significantly.

Training	BEIR-subset (ndcg@10)	SciDocs (avg.)	ICLR2022 (ndcg@10)
ICrop.	0.244	74.93	0.370
MS MARCO	0.270	68.72	0.260
ICrop. + MS MARCO	**0.324**	**75.09**	**0.396**
SPECTER	0.207	**79.75**	0.407
ICrop. + SPECTER	**0.248**	79.40	**0.471**
unarXiv doc2doc	0.170	75.42	0.378
ICrop. + unarXiv doc2doc	**0.251**	**78.19**	**0.467**
unarXiv context2doc	0.252	75.08	0.379
ICrop. + unarXiv context2doc	**0.285**	**78.29**	**0.450**

Table 3: Comparison of training on a single objective versus using independent croping (ICrop) and finetuning on the target objective (Ind. Crop. + {target}). BEIR subset is the average ndcg of the 5 tasks, SciDocs avg is the average metric of all the tasks. Results in bold are the best result given a target objective.

Independent cropping. Furthermore, we study how well self-supervised pre-training on an information retrieval task performs. In this experiment, we train a model using independent cropping and finetune it on four target objectives. Table 3 presents the results when finetuning on MS MARCO , SPECTER, unarXiv document to document, and unarXiv context to document objectives. We find that independent cropping is an effective pre-training method. For every one of the target objectives (except SPECTER on the SciDocs benchmark), pre-training using the self-supervised method leads to an increase in performance compared to only training on the target objective. The results are consistent across the three evaluation metrics for all the target objectives except SPECTER (the performance on SciDocs is slightly less but similar).

Using citation contexts as semi-supervised signal. In addition, we make use of citation contexts to extract semi-supervised relevance signals. In particular, we study two ways to define relevancy: the first is co-occurring documents within a citation context (*unarXiv doc2doc*), and the second uses the context as the query for any document appearing in it (*unarXiv context2doc*). The last four rows in Table 3 presents the doc2doc in context vs. context2doc. Using unarXiv doc2doc as a single training objective does not perform well across the three evaluation metrics but using it as a target objective after training on independent cropping improves the performance but does not compare to finetuning on SPECTER which gets similar results on BEIR but higher results on SciDocs and ICLR2022 . One explanation could be the difference in data set size: SPECTER training set contains 684,100 triplets and unarXiv doc2doc only contains 456,766. Using citation context as queries (unarXiv context2doc) is a better alternative. When pre-trained on independent cropping and finetuned on unarXiv context2doc, our model performs well on the three evaluation metrics. In particular, the model performs well on the subset of BEIR whose tasks contain mostly short queries. We find that using citation contexts, which are shorter than documents (on average 29.8 words per context), is an effective way to introduce query to document supervision for the scientific document domain.

Pre-train data	BEIR	SciDocs	ICLR 2022
Baselines			
MSMARCO	0.270	68.72	0.260
SPECTER	0.207	79.75	0.407
ICrop.	0.244	74.93	0.370
Finetuning on InPars			
MSMARCO	0.300	69.70	0.248
SPECTER	0.304	74.88	0.300
ICrop	0.313	75.12	0.341

Table 4: Finetuning models on InPars synthetic data. The first 3 rows are models trained on a single objective. Metrics are ndcg@10 for BEIR and ICLR2022 , and average over all tasks for SciDocs.

Using synthetic data. Finally, following (Bonifacio et al., 2022), we experimented with InPars, the introduction of in domain synthetic data (query document pairs generated using GPT-3 (Brown et al., 2020)). Table 4 presents the results of finetuning on InPars data compared to single objective training. We find that InPars significantly improves the performance on BEIR regardless of the pre-training objective. The performance on SciDocs increases when for both MS MARCO and independent cropping pre-training. The results are mitigated when finetuning a model trained on SPECTER, the performance on BEIR increases by 50% with the cost of 6.1% and 11.1% decrease on SciDocs and ICLR2022 respectively. The synthetic data contains mostly short queries, therefore the increase in performance on BEIR, which contains a majority of short query to document tasks, is expected. Future work could explore generating long query synthetic data.

6 Conclusion

We explored multi-objective dense retrieval training as a way to optimize models for both short queries and document queries. We study three ways to combine objectives (in-batch mixing, alternating between batches, and finetuning). We find that using multiple objectives is a way to train dense retrievers that perform well for short and long query retrieval. Considering the performances across a subset of BEIR, SciDocs and ICLR2022 our best models achieve an average relative improvement between 13.4 and 19.2% compared to the best single objective models. Our work here focused on bi-encoders, and future work could explore whether multi-objective training is also beneficial for a cross-encoder architecture (Lin et al., 2021).

Furthermore, we find that pre-training a model using independent cropping and finetuning it on a target objective consistently improves the retrieval performance compared to only training on the target objective.

We also introduced context-to-document, a new weakly supervised training objective using the citation context sentence as the query for the cited document. This signal outperforms the co-cited relevance signal and improves the model performance on short query retrieval. For future work, we would like to explore how to pre-filter citation contexts that do not contain useful information to identify a relevant document, thus removing potential noise from the training data.

Finally, we released a new document to document retrieval data set composed of ICLR 2022 papers and 1,844 human relevance judgement.

All our experiments were conducted using a MiniLM-L6 (22.7 million parameters) model which is more than 10 times smaller than BERT (Devlin et al., 2019) (345 million parameters). Future work should consider how multi-objective training scales to larger models.

With this work, we hope to make a step in the direction of a multi-purpose document representation to reduce storage space, computational resources, and increase the operational efficiency of scientific retrieval systems.

Acknowledgements

We thank Andrew Yates, Assistant Professor in the Information Retrieval Lab at the University of Amsterdam, for assistance and feedback with the methodology and experiments.

References

Amin Abolghasemi, Suzan Verberne, and Leif Azzopardi. 2022. Improving bert-based query-by-document retrieval with multi-task optimization.

Sophia Althammer, Sebastian Hofstätter, Mete Sertkan, Suzan Verberne, and Allan Hanbury. 2022. Parm: A paragraph aggregation retrieval model for dense document-to-document retrieval.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. *CoRR*, abs/1805.02262.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack

Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Rich Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Giovanni Colavizza, Kevin W. Boyack, Nees Jan van Eck, and Ludo Waltman. 2017. The closer the better: Similarity of publication pairs at different co-citation levels.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Marzieh Fadaee, Olga Gureenkova, Fernando Rejon Barrera, Carsten Schnober, Wouter Weerkamp, and Jakub Zavrel. 2020. A new neural search and insights platform for navigating and organizing ai research.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity.

Malte Ostendorff, Till Blume, Terry Ruas, Bela Gipp, and Georg Rehm. 2022a. Specialized document embeddings for aspect-based similarity of research papers.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022b. Neighborhood contrastive learning for scientific document representations with citation embeddings.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tarek Saier and Michael Färber. 2020. unarXive: A Large Scholarly Data Set with Publications' Full-Text, Annotated In-Text Citations, and Links to Metadata. *Scientometrics*, 125(3):3085–3108.

Shicheng Tan, Shu Zhao, and Yanping Zhang. 2022. Coherence-based distributed document representation learning for scientific documents.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: constructing a pandemic information retrieval test collection. *CoRR*, abs/2005.04474.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *EMNLP*.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Multi-view document representation learning for open-domain dense retrieval.