Unsupervised Dense Retrieval Training with Web Anchors

Yiqing Xie
Carnegie Mellon University
USA
yiqingxi@andrew.cmu.edu

Xiao Liu
Microsoft
USA
xiliu2@microsoft.com

Chenyan Xiong
Microsoft
USA
chenyan.xiong@microsoft.com

ABSTRACT

In this work, we present an unsupervised retrieval method with contrastive learning on web anchors. The anchor text describes the content that is referenced from the linked page. This shows similarities to search queries that aim to retrieve pertinent information from relevant documents. Based on their commonalities, we train an unsupervised dense retriever, Anchor-DR, with a contrastive learning task that matches the anchor text and the linked document. To filter out uninformative anchors (such as “homepage” or other functional anchors), we present a novel filtering technique to only select anchors that contain similar types of information as search queries. Experiments show that Anchor-DR outperforms state-of-the-art methods on unsupervised dense retrieval by a large margin (e.g., by 5.3% NDCG@10 on MSMARCO). The gain of our method is especially significant for search and question answering tasks. Our analysis further reveals that the pattern of anchor-document pairs is similar to that of search query-document pairs.[1]

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**.

KEYWORDS

Dense retrieval, Unsupervised dense retrieval, Contrastive learning

ACM Reference Format:
Yiqing Xie, Xiao Liu, and Chenyan Xiong. 2023. Unsupervised Dense Retrieval Training with Web Anchors. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592080

1 INTRODUCTION

Dense retrieval matches queries and documents in the embedding space [15, 16, 26], which can capture the semantic meaning of the text and handle more complex queries compared to traditional sparse retrieval methods [23]. Due to the scarcity of labeled data in certain domains, including legal and medical, numerous recent studies have focused on unsupervised dense retrieval, which trains dense retrievers without annotations [11, 13, 14, 18].

One of the most common approaches of unsupervised dense retrieval is to design a contrastive learning task that approximates retrieval [3, 11, 13, 14, 18, 19, 22], yet it is nontrivial to construct contrastive pairs. Most existing methods construct contrastive pairs from the same context, such as a sentence and its context [14], or two individual text spans in a document [11, 13, 19]. The relation between these co-document pairs is different from query-document pairs in search or question answering, where the query aims to seek information from the document. LinkBERT [27] leverage text spans sampled from a pair of linked Wikipedia pages. However, such text spans are not guaranteed to have high relevance. Few other methods train a model to generate queries from documents [2, 18], but they either require large language models or huge amounts of training data.

In this work, we present Anchor-DR, an unsupervised dense retriever that is trained on predicting the linked document of an anchor given its anchor text. The text on the anchor of hyperlinks typically contains descriptive information that the source document cites from the linked document, suggesting that the anchor-document pairs exhibit resemblances to query-document pairs in search, where the search query describes the information that the user is required from the relevant document. As a result, we present to train Anchor-DR to match the anchor text and its linked document with a contrastive objective.

Although the relation between anchor-document pairs is typically similar to that of search queries and relevant documents, there also exist a large number of uninformative anchors. For example, a web document may use anchor links to redirect to the linked document (e.g., “homepage” or “website”). Such anchor-document pairs do not resemble the relation between search queries and documents and may introduce noise to our model. We thus design a few heuristic rules to filter out functional anchors, such as headers/footers or anchors in the same domain. In addition, we train a classifier with a small number of high-quality search queries to further identify anchors containing similar types of information as real search queries.

Experiment results show that Anchor-DR outperforms state-of-the-art unsupervised dense retrievers by a large margin on two widely adopted retrieval datasets, MSMARCO [1] and BEIR [24] (e.g., by 5.3% NDCG@10 on MSMARCO). The improvement of Anchor-DR is most significant on search and question answering tasks, suggesting that compared to the contextual relation between co-document text spans [11, 13], the referral relation between anchor-document pairs is more similar to the information-seeking relation between search query-document pairs. We further present examples to show that anchor-document pairs indeed have similar patterns as query-document pairs.

2 RELATED WORK

Dense Retrieval. Dense retrieval is the technique of using dense vector representations of text to retrieve relevant documents [5, 12].

[1] Code available at https://github.com/Veronicium/AnchorDR.

SIGIR ’23, July 23–27, 2023, Taipei, Taiwan, China.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3592080

With the development of pretrained language models [6, 15], recent works have developed various techniques for dense retrieval, including retrieval-oriented pretraining [11, 13, 19] and negative selection [26]. While dense retrieval has exhibited remarkable effectiveness in contrast to traditional sparse retrieval approaches [23], its benefits are generally confined to supervised settings that involve an adequate amount of human annotations [24].

Unsupervised dense retrieval. Previous work on unsupervised dense retrieval mainly adopts contrastive learning to model training. ICT [14] matches the surrounding context of a random sentence. SPAR [3] uses random sentences as queries with positive and negative passages ranked by the BM25 score. Co-condenser [11], COCO-LM [19], and contriever [13] regard independent text spans in one document as positive pairs. QExt [18] further improves their work by selecting the text span with the highest relevance computed by an existing pretrained model. A few other research works use neural models to generate queries, such as question-like queries [2] or the topic, title, and summary of the document [18]. However, both works require a large-scale generation system.

Leveraging web anchors in retrieval. Web anchors have been widely applied to classic approaches for information retrieval [4, 7, 8, 10, 28]. Recently, HARP [17] designs several pretraining objectives leveraging anchor texts, including representative query prediction or query disambiguation modeling. ReInfoSelect [29] learns to select anchor-document pairs that best weakly supervise the neural ranker. However, these methods either focus on classic bag-of-word modeling or apply a cross-encoder architecture that does not fit the setting of dense retrieval.

3 METHODOLOGY

We present an unsupervised dense retrieval method that trains the model to match the representations of anchor text and its linked document. This section describes the contrastive learning task of anchor-document prediction and the anchor filtering process.

3.1 Contrastive Learning with Anchor-Document Pairs

Based on the commonalities between anchor-document pairs and query-document pairs [4, 7, 8, 10, 28], we compute the representation of each anchor and document with our model, Anchor-DR, and trains it with a contrastive objective of matching anchor text and its linked document:

$$\mathcal{L}(a, d_+) = -\frac{\exp(sim(a, d_+))}{\exp(sim(a, d_+)) + \sum_{d_- \in Neg(a)} \exp(sim(a, d_-))} \quad (1)$$

$$sim(a, d) = \langle f_\theta(a), f_\theta(d) \rangle, \quad (2)$$

where f_θ is our presented model, Anchor-DR, with T5 [21] as its backbone, the sequence embedding $f_\theta()$ is the embedding of the first token output by the decoder of Anchor-DR, (a, d_+) is the anchor text and its linked document, and $Neg(a)$ is the set of negative documents sampled from the whole dataset. In practice, we use BM25 negatives in the first iteration [15] and use the negatives mined by Anchor-DR in the following iterations [26].

In inference, we feed the query and all the documents into Anchor-DR separately and use the embedding of the first token in the decoder output as the sequence embedding. Then we rank all the documents by their similarity to the query: $sim(q, d) = \langle f_\theta(q), f_\theta(d) \rangle$, where f_θ denotes Anchor-DR.

Table 1: The statistics of ClueWeb22 anchor training data.

# of docs			# of anchors		
Raw	After filt. by rules	After filt. by model	Raw	After filt. by rules	After filt. by model
60.49M	10.17M	3.97M	117.11M	20.66M	4.25M

3.2 Anchor Filtering

While some anchor-document pairs exhibit strong similarities with query-document pairs in search, others do not. For instance, "homepage" or "website" and their linked documents hold entirely distinct relations with query-document pairs. Including these pairs in the training data may introduce noise to our model. As a result, we first apply a few heuristic rules and then train a lightweight classifier to filter out uninformative anchor text.

Anchor filtering with heuristic rules. We observe that a large number of uninformative anchors are functional anchors and these anchors mainly exist between pages within the same website. Consequently, we filter out anchor text that falls in the following categories: (1) *In-domain anchors*, where the source and target page share the same domain; (2) *Headers or footers*, which are detected by specific HTML tags, such as <header> and <footer>; and (3) *Keywords indicating functionalities*, which are manually selected from anchors with top 500 frequency. [2]

Anchor filtering with query classifier. We train a lightweight query classifier to learn the types of information that is typically contained in search queries about relevant documents. Specifically, we use the *ad-hoc* queries provided by WebTrack [9] as positive examples. These small number of queries are manually selected to reflect important characteristics of authentic Web search queries for each year. As for negative examples, we sample a subset of anchors before filtering by our rules, which has the same size as positive examples We train the query classifier with the Cross-Entropy Loss:

$$\mathcal{L} = \sum_x \mathbf{1}_{Pos} \cdot log(g(x)) + \mathbf{1}_{Neg} \cdot log(1 - g(x)), \quad (3)$$

where g is a miniBERT-based [25] model. After training the query classifier, we rank all the anchor text by the logits of the positive class (i.e., similarity to search queries) and only keep the top 25%.

4 EXPERIMENTS

In this section, we describe the experiment setups, compare Anchor-DR with baselines and ablations, and analyze its effectiveness.

4.1 Experimental Setup

We evaluate Anchor-DR on two public datasets: MSMARCO [1] and BEIR [24] for unsupervised retrieval, where we directly apply the methods to encode test queries and documents without supervision. We report the nDCG@10 results following previous works [13, 18].

Training data. We train Anchor-DR on a subset of the ClueWeb22 dataset [20]. To preprocess the data, we first randomly sampled a subset of English documents with at least one in-link. After that,

[2] We list the keywords in github.com/Veronicium/AnchorDR/blob/main/anchor_filtering

Table 2: Unsupervised retrieval results on MSMARCO and BEIR under nDCG@10. The best result for each task is marked in bold. The best result among dense retrievers is underlined. We follow previous work [13] and report the average performance on 14 BEIR tasks and MSMARCO (BEIR14+MM). The results of coCondenser and results with † are evaluated using their released checkpoints. The results of other baselines are copied from their original papers.

Model (→)	BM25	coCondenser	Contriever	SPAR	QExt	**Anchor-DR**
Training Data	-	MSMARCO	Wiki+CCNet	Wiki	Pile-CC	ClueWeb
# Training Pairs	-	8.8M	3M+707M	22.6M	52.4M	4.25M
MS MARCO	22.8	16.2	20.6	19.3	20.6	**25.9**
TREC-COVID	65.6	40.4	27.4	53.1	53.5	**77.4**
BioASQ	**46.5**	22.7	32.7†	-	-	31.9
NFCorpus	**32.5**	28.9	31.7	26.4	30.3	30.8
NQ	32.9	17.8	25.4	26.2	27.2	**33.6**
HotpotQA	**60.3**	34.0	48.1	57.2	47.9	53.2
FiQA-2018	23.6	**25.1**	24.5	18.5	22.3	21.1
Signal-1M	**33.0**	21.4	25.0†	-	-	20.9
TREC-NEWS	39.8	25.4	35.2†	-	-	**45.5**
Robust04	**40.8**	29.8	32.7†	-	-	40.1
ArguAna	31.5	**44.4**	37.9	42.0	39.1	29.1
Touchè-2020	**36.7**	11.7	19.3	26.1	21.6	25.0
CQADupStack	29.9	**30.9**	28.4	27.9	27.1	29.1
Quora	78.9	82.1	**83.5**	70.4	82.7	72.1
DBPedia-ent	31.3	21.5	29.2	28.1	29.0	**34.1**
SCIDOCS	15.8	13.6	14.9	13.4	14.7	**15.9**
FEVER	**75.3**	61.5	68.2	56.9	59.7	71.1
Climate-fever	**21.3**	16.9	15.5	16.4	17.7	20.6
SciFact	**66.5**	56.1	64.9	62.6	64.4	59.4
BEIR14+MM	**41.7**	33.4	36.0	36.2	37.0	39.9
All Avg.	**41.3**	31.6	35.0	-	-	38.8
Best on	**9**	3	1	0	0	6

we use rules and then train a query classifier to filter out uninformative anchors, as introduced in Sec. 3.2. Finally, we sample at most 5 in-links for each document. The statistics of the anchors and documents after each step of filtering are shown in Table 1. Note that ClueWeb22 has in total of 52.7B anchors, hence we are able to further scale up our model in the future.

Implementation details. For continuous pretraining on anchor-document prediction, we train our model with BM25 negatives for one epoch and with ANCE negatives [26] for another epoch. We use a learning rate of 1e-5 and a batch size of 128 positive pairs. The query classifier is trained on the *adhoc* test queries of WebTrack 2009 - 2014 [9], which contains 300 queries in total.

Baselines. We compare Anchor-DR with a sparse retrieval method: BM25 [23] and four unsupervised dense retrieval methods: coCondenser [11], Contriever [13], SPAR Λ (trained on Wikipedia) [3], and QExt-PLM (trained on Pile-CC with MoCo) [18]. All these dense retrieval methods construct contrastive pairs in an unsupervised way: either by rules [11, 13], lexical features [3], or with pretrained models [18]. Note that we do not compare with methods that require large-scale generation system to generate contrastive pairs, such as QGen [18] or InPars [2], as their generators either require additional human annotations or have significantly larger sizes compared to our model (e.g., 6B vs. 220M).

As for ablation studies, we substitute the anchor-document prediction task with two other contrastive tasks: *ICT* [14], which considers a document and a sentence randomly selected from the document as positive pairs, and *co-doc* [11], which treats two text sequences from the same document as positive pairs. We also compare to *Anchor (rule only)*, which removes the query classifier and only uses rules to filter anchors. For a fair comparison, we train all the ablations on the same subset of documents in ClueWeb22.

Table 3: nDCG@10 of models trained with different contrastive tasks on the same subset of documents, with 400K documents and 400K contrastive pairs. T-test shows Anchor-DR outperforms co-doc on All Avg. with p-value < 0.05.

Model (→)	ICT	co-doc	Anchor (rule only)	**Anchor-DR**
MSMARCO	20.9	19.9	20.3	**22.1**
TREC-COVID	65.0	64.4	**72.4**	70.6
BioASQ	29.7	26.7	29.7	**30.8**
NFCorpus	27.1	24.0	24.7	**28.4**
NQ	23.4	27.9	30.7	**31.0**
HotpotQA	39.8	38.9	41.5	**48.9**
FiQA-2018	**20.4**	17.8	19.1	18.2
Signal-1M	19.7	18.1	19.5	**21.1**
TREC-NEWS	37.1	39.2	**43.3**	42.5
Robust04	30.4	34.6	34.9	**38.2**
ArguAna	39.7	**45.1**	26.0	26.5
Touchè-2020	23.0	25.4	**27.4**	25.2
CQADupStack	26.3	26.2	**26.7**	24.8
Quora	76.6	74.9	**77.3**	71.6
DBPedia-ent	25.2	26.7	27.5	**31.4**
SCIDOCS	14.0	13.6	13.8	**14.7**
FEVER	57.7	56.5	**72.2**	69.8
Climate-fever	19.5	**20.0**	21.2	18.3
SciFact	54.1	54.3	50.4	**56.1**
BEIR14 + MM	35.5	35.7	36.7	**37.2**
All Avg.	34.2	34.4	35.7	**36.3**

4.2 Main Results

Table 2 shows the unsupervised retrieval results on MSMARCO and BEIR. Anchor-DR outperforms all the dense retrieval baselines on MSMARCO and BEIR with a large margin (e.g., by 2.9% nDCG@10 on BEIR14+MM and 3.8% on all datasets). Furthermore, compared to other dense retrievers, Anchor-DR achieves the best performances across a majority of datasets. indicating that our method can be generalized to a wide range of domains and retrieval tasks.

We observe that Anchor-DR exhibits strong performance in specific subsets of tasks. For instance, Anchor-DR achieves a large performance gain of 11.8% nDCG@10 on TREC-COVID, but it is outperformed by other baseline methods on ArguAna and Quora.

4.3 Ablation Study

To demonstrate the effectiveness of our anchor-doc prediction task, we perform ablation studies in Table 3. We observe that Anchor-DR outperforms both methods. Additionally, *ICT* and *co-doc* have less than 1% performance gap on 7 out of 19 datasets. This is probably because the contrastive learning pairs in both methods contain contextual information about each other. Anchor-DR also outperforms *Anchor (rule only)*, indicating that it is effective to train on anchor texts with higher similarities to search queries.

4.4 Performance Analysis

Performance breakdown. The results in Table 2 show that Anchor-DR achieves strong performance in a majority of datasets but

Table 4: Examples of the query-document pairs in two BEIR datasets: ArguAna and TREC-COVID, the co-document text pairs (co-doc), and the anchor-document pairs (Anchor-DR).

***Dataset**: ArguAna* **Query**: Becoming a vegetarian is an environmentally friendly thing to do. Modern farming is one of the main sources of pollution in our rivers, and as long as people continue to buy fast food ... **Document**: Health general weight philosophy ethics You don't have to be vegetarian to be green. Many special environments have been created by livestock farming, for example chalk down land in England and mountain pastures ...

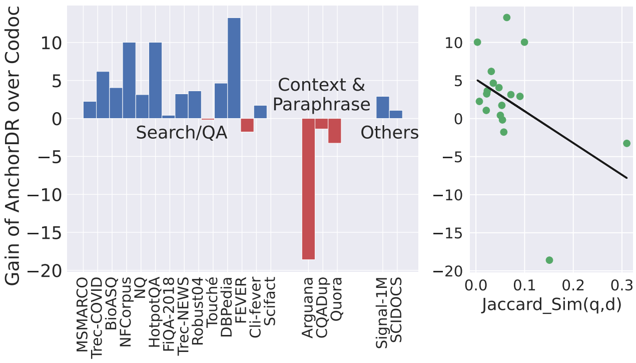
***Dataset**: TREC-COVID* **Query**: what causes death from Covid-19? **Document**: Predicting the ultimate outcome of the COVID-19 outbreak in Italy: During the COVID-19 outbreak, it is essential to monitor the effectiveness of measures taken by governments on the course of the epidemic. Here we show that there is already a sufficient amount of data collected in Italy to predict the outcome of the process ...

***Method**: Codoc* **Query #1**: Going vegetarian is one of the best things you can do for your health. **Document #1**: We publish a quarterly magazine The Irish Vegetarian, with features and our roundup of news and events of interest to Irish vegetarians. Get involved! There are lots of ways to get involved. You can read our Going Vegetarian page. You can pick up a copy of The Irish Vegetarian. You can come to a Meetup meeting ...

Query #2: COVID-19 vaccines designed to elicit neutralizing antibodies may sensitize vaccine recipients to severe diseases **Document #2**: According to a study that examined how informed consent is given to COVID-19 vaccinetrial participants, disclosure forms fail to inform volunteers that the vaccine might make them susceptible to more severe disease. The study, "Informed Consent Disclosure to Vaccine Trial Subjects of Risk of COVID-19 Vaccine ...

***Method**: Anchor-DR* **Query #1**: Vegetarian Society of Ireland **Document #1**: The Vegetarian Society of Ireland is a registered charity. Our aim is to increase awareness of vegetarianism in relation to health, animal welfare and environmental perspectives. We support both vegetarian and vegan aims. Going vegetarian is one of the best things you can do for your health, for animals and for the planet ...

Query #2: How COVID19 Vaccine Can Destroy Your Immune System **Document #2**: According to a study that examined how informed consent is given to COVID-19 vaccine trial participants, disclosure forms fail to inform volunteers that the vaccine might make them susceptible to more severe diseases...



(a) Task Type **(b) Lexical Overlap**

Figure 1: Performance gain of Anchor-DR over codoc on different datasets under nDCG@10.

not in others. To analyze the effectiveness of Anchor-DR on different datasets, we categorize the datasets into three subsets: (1) Search/QA, where the query is a question or keywords related to the document; (2) Context/Paraphrase, where the query and document contain coherent or overlapping information; and (3) Others. Figure 1(a) shows that Anchor-DR performs better on Search/QA datasets and co-doc is better on Context/Paraphrase datasets. The results are consistent with our hypothesis that the referral relation between query-document pairs is similar to the information-seeking relation between search queries and relevant documents.

We further quantitatively analyze the information pattern of query-document pairs captured by Anchor-DR and co-doc. Figure 1(b) shows the performance gap between Anchor-DR and co-doc versus the degree of information overlap between queries and documents in each test dataset, which is measured using Jaccard Similarity. We observe that Anchor-DR performs much better on datasets where queries and documents contain less overlapping information. The primary emphasis of datasets with high query-document similarity is mainly on paraphrasing and coherency, which are distinct from the relation between search queries and documents.

Case studies. We show in Table 4 the contrastive pairs of Anchor-DR and co-doc, as well as the positive pairs in ArguAna and TREC-COVID, which represent the Search/QA and Context/Paraphrase datasets. The query-doc pairs of ArguAna are arguments around the same topic, which are coherent and have similar formats. Similarly, the contrastive pairs of co-doc contain either coherent (e.g., the claim and recent work of the vegetarian society) or repeating information (e.g., COVID vaccine may cause diseases), which may explain its good performance on Context/Paraphrase datasets.

In contrast, in TREC-COVID, the answer to the query is contained in the document. As shown in Table 4, the anchor text in Anchor-DR could be the topic of the linked document, or in the format of a question. In both examples, the anchor text can serve as a search query and the document can provide the information the query is seeking, which could be the reason why Anchor-DR achieves strong performance on the Search/QA datasets.

5 CONCLUSION

We train an unsupervised dense retrieval model, Anchor-DR, leveraging the rich web anchors. In particular, we design a contrastive learning task: anchor-document prediction to continuously pretrain Anchor-DR. Additionally, we apply predefined rules and train a query classifier to filter out uninformative anchors. Experiments on two public datasets: MSMARCO and BEIR show that Anchor-DR significantly outperforms the state-of-the-art dense retrievers on unsupervised retrieval. Our analyses provide a further comparison of the patterns of information contained in our contrastive learning pairs and query-document pairs in test datasets.

REFERENCES

[1] 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. https://arxiv.org/abs/1611.09268

[2] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Unsupervised Dataset Generation for Information Retrieval *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2387–2392. https://doi.org/10.1145/3477495.3531863

[3] Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2022. Salient Phrase Aware Dense Retrieval: Can a Dense Retriever Imitate a Sparse One?. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 250–262. https://aclanthology.org/2022.findings-emnlp.19

[4] Na Dai and Brian D. Davison. 2010. Mining Anchor Text Trends for Retrieval. In *Advances in Information Retrieval*, Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan Rüger, and Keith van Rijsbergen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 127–139.

[5] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* 41 (1990), 391–407.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics. https://aclanthology.org/N19-1423

[7] Zhicheng Dou, Ruihua Song, Jian-Yun Nie, and Ji-Rong Wen. 2009. Using Anchor Texts with Their Hyperlink Structure for Web Search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 227–234. https://doi.org/10.1145/1571941.1571982

[8] Nadav Eiron and Kevin S. McCurley. 2003. Analysis of Anchor Text for Web Search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (Toronto, Canada) *(SIGIR '03)*. Association for Computing Machinery, New York, NY, USA, 459–460. https://doi.org/10.1145/860435.860550

[9] Junlan Feng, Valerie Torres, Daniel Sheleheda, and Cynthia Cama. 2008. Web-Track: Mining and Tracking Content and Structure of Websites. *Proceedings of the 2008 International Conference on Data Mining, DMIN 2008*, 667–673.

[10] Raya Fidel, Harry Bruce, Annelise Mark Pejtersen, Susan Dumais, Jonathan Grudin, and Steven Poltrock. 2000. Collaborative Information Retrieval (CIR). *The New Review of Information Behaviour Research* 1 (January 2000), 235–247. https://www.microsoft.com/en-us/research/publication/collaborative-information-retrieval-cir/

[11] Luyu Gao and Jamie Callan. 2022. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 2843–2853. https://doi.org/10.18653/v1/2022.acl-long.203

[12] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 2333–2338. https://doi.org/10.1145/2505515.2505665

[13] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised Dense Information Retrieval with Contrastive Learning. https://doi.org/10.48550/ARXIV.2112.09118

[14] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=jKN1pXi7b0

[15] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online. https://aclanthology.org/2020.emnlp-main.550

[16] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy. https://aclanthology.org/P19-1612

[17] Zhengyi Ma, Zhicheng Dou, Wei Xu, Xinyu Zhang, Hao Jiang, Zhao Cao, and Ji-Rong Wen. 2021. Pre-Training for Ad-Hoc Retrieval: Hyperlink is Also You Need. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 1212–1221. https://doi.org/10.1145/3459637.3482286

[18] Rui Meng, Ye Liu, Semih Yavuz, Divyansh Agarwal, Lifu Tu, Ning Yu, Jianguo Zhang, Meghana Bhat, and Yingbo Zhou. 2022. Unsupervised Dense Retrieval Deserves Better Positive Pairs: Scalable Augmentation with Query Extraction and Generation. https://doi.org/10.48550/ARXIV.2212.08841

[19] Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and contrasting text sequences for language model pretraining. In *Conference on Neural Information Processing Systems*.

[20] Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Visual and Semantic Information. https://doi.org/10.48550/ARXIV.2211.15848

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. https://doi.org/10.48550/ARXIV.1910.10683

[22] Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2021. Learning to Retrieve Passages without Supervision. https://doi.org/10.48550/ARXIV.2112.07708

[23] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. (2009). https://doi.org/10.1561/1500000019

[24] 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)*. https://openreview.net/forum?id=wCu6T5xFjeJ

[25] Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and Practical BERT Models for Sequence Labeling. https://arxiv.org/abs/1909.00100

[26] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations*. https://openreview.net/forum?id=zeFrfgyZln

[27] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining Language Models with Document Links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.551

[28] Xing Yi and James Allan. 2010. A Content Based Approach for Discovering Missing Anchor Text for Web Search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 427–434. https://doi.org/10.1145/1835449.1835521

[29] Kaitao Zhang, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2020. Selective Weak Supervision for Neural Information Retrieval. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 474–485. https://doi.org/10.1145/3366423.3380131