Referral Augmentation for Zero-Shot Information Retrieval

Michael Tang **Shunyu Yao** **John Yang** **Karthik Narasimhan**
Department of Computer Science, Princeton University
{mwtang, shunyuy, jy1682, karthikn}@princeton.edu

Abstract

We propose Referral-Augmented Retrieval (RAR), a simple technique that concatenates document indices with *referrals*: text from other documents that cite or link to the given document. We find that RAR provides significant performance gains for tasks across paper retrieval, entity retrieval, and open-domain question-answering in both zero-shot and in-domain (e.g., fine-tuned) settings. We examine how RAR provides especially strong improvements on more structured tasks, and can greatly outperform generative text expansion techniques such as DocT5Query (Nogueira et al., 2019) and Query2Doc (Wang et al., 2023), with a 37% and 21% absolute improvement on ACL paper retrieval, respectively. We also compare three ways to aggregate referrals for RAR. Overall, we believe RAR can help revive and re-contextualize the classic information retrieval idea of using anchor texts to improve the representations of documents in a wide variety of corpuses in the age of neural retrieval.[1]

1 Introduction

Zero-shot information retrieval, a task in which both test queries and corpora are inaccessible at training time, closely mimics real-world deployment settings where the distribution of text changes over time and the system needs to continually adapt to new queries and documents. Prior work (Thakur et al., 2021) finds that without access to training on in-domain query-document pairs or task-specific document relations, most dense models dramatically underperform simple sparse models like BM25, pointing to poor generalization. At the same time, sparse models struggle to reconcile different surface forms, leading to the so-called *lexical gap* between queries and documents in different tasks.

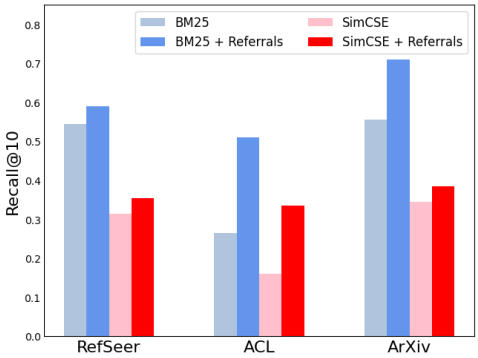

Figure 1: Referral-Augmented Retrieval (RAR) improves zero-shot document retrieval across a variety of models and datasets, such as paper retrieval, shown here.

While the zero-shot setting lacks query-document pairs, our key insight is to leverage inter-document relations that provide multiple views of the same information to provide a more comprehensive representations of the concepts in a document. We propose **Referral-Augmented Retrieval (RAR)**, a simple technique that augments the text of each document in a retrieval index with passages from other documents that contain citations or hyperlinks to it. This use of inter-document information is reminiscent of Google's BackRub and PageRank algorithms.

In the age of pretrained models, we revisit this classical intuition on new, dense retrievers such as SimCSE and DPR (Gao et al., 2021; Karpukhin et al., 2020), as well as new domains with referral links like the Semantic Scholar citation graph (Lo et al., 2020) and Wikipedia entity graph (Hasibi et al., 2017) (more on the recontextualization in Appendix Section C).

[1]Code: https://github.com/michaelwilliamtang/referral-augment

Findings of the Association for Computational Linguistics ACL 2024, pages 13452–13461

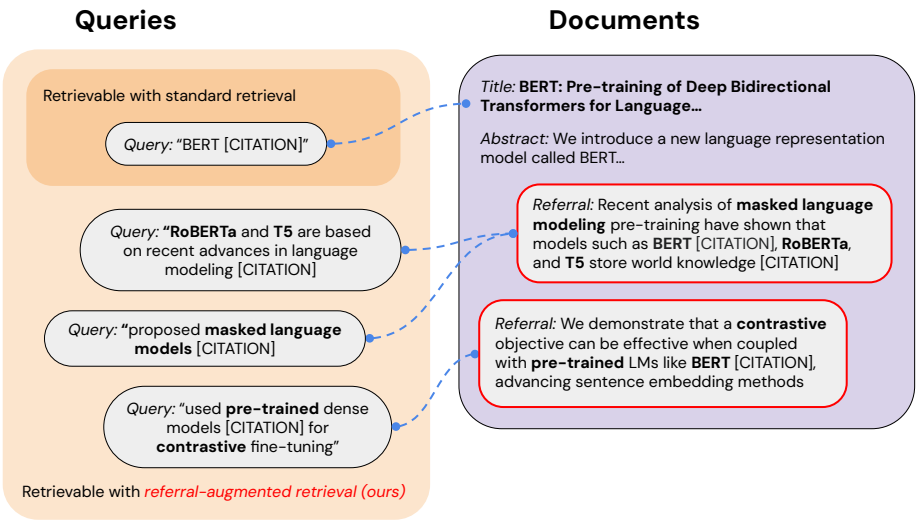

Figure 2: Illustration of RAR: by augmenting the index with information from documents that refer to the original document (in red), we correctly retrieve the target document for a wider range of possible queries compared to standard methods.

We evaluate RAR across a broad range of domains and corpora, and find that it significantly boosts zero-shot retrieval performance on **five out of six zero-shot IR tasks** and **greatly outperforms generative text expansion techniques** such as DocT5Query (Nogueira et al., 2019) and Query2Doc (Wang et al., 2023). We **probe the strengths and limitations of referral augmentation**: we find that it struggles on less structured and particularly open-ended tasks like multihop question answering, but on most other tasks it can help dramatically improve performance in both zero-shot and in-domain settings and for both sparse and dense models. Since augmentation is simple, occurs entirely at indexing time, and requires no expensive model inference, it provides a training-free method to adapt retrievers to new areas with relatively structured queries and allows the continuous addition of new documents.

2 Related Work

Zero-shot information retrieval Following the popularization of the zero-shot information retrieval setting across domains by the BEIR (Thakur et al., 2021) benchmark, many methods have been proposed, including leveraging large language models as zero-shot rerankers using document-to-question continuation probabilities (Sachan et al., 2022) and zero-shot retrievers by writing query augmentations based on preliminary retrieved candidates (Shen et al., 2023). These are largely complementary to our work, which draws on underutilized metadata instead of the memorized knowledge of expensive generative models.

Query and document expansion Query expansion and document expansion techniques such as DocT5Query, HyDE, and Query2Doc (Nogueira et al., 2019; Gao et al., 2022; Wang et al., 2023) were proposed to decrease the lexical gap between queries and documents, typically using generative models to rephrase and add context. We discuss a unifying view of RAR as expansion in Section B and empirically find that RAR outperforms generative expansion based on both specialized models and large language models like GPT-3 in Section 4.3.

Hyperlink anchor texts for web retrieval One classic line of work explores the utility of hyperlink anchor text in improving site discovery for search engines. McBryan, Brin and Page, and Kleinberg's seminal papers on internet search systems mention using incoming links as a marker of a given page's relevance as well as storing the linking anchor text as metadata (McBryan, 1994; Brin and Page, 1998; Kleinberg, 1999); Craswell and Hawking implement a site retriever using BM25 on this metadata, combining incoming anchor texts into an

"anchor document" (Craswell et al., 2001), and this method is later refined for web search tasks using different ad hoc anchor and content-based rankings (Westerveld et al., 2001; Koolen and Kamps, 2010; Dou et al., 2009). Twenty years after these influential works from classical IR, we generalize the idea of referral augmentation in a model-agnostic (e.g., both sparse and dense retrieval) and domain-agnostic (e.g., ACL, ArXiv, Wikipedia) way. Further, while traditional anchor texts are formatted as a few words without corresponding context, RAR can leverage the full sentence- or passage-level context containing the referral as a semantic augmentation, which better suits modern neural IR approaches (e.g. SimCSE sentence embedding) with stronger semantic understanding.

Hyperlinks and citations for contrastive training Another line of work explores using hyperlinks and citations for *training*, using referrals as pseudo-queries. Entity retrieval models (Mitra et al., 2017; Wu et al., 2022) explore pre-training using the anchor-text-document pairs, whereas paper retriever models (Gu et al., 2022; Cohan et al., 2020) fine-tune using citing-cited paper pairs. In contrast, we focus on using hyperlinks and citations to build *training-free* document augmentations that work with any off-the-shelf encoder. We also empirically find in Table 2 that models trained in this way still benefit from RAR.

3 Method

3.1 Preliminaries

Formally, given a set of queries Q and documents D, retrieval can be described as the task of learning a similarity function $\text{sim}\,(q, d)$ between a query $q \in Q$ and document $d \in D$, where top-k retrieval is equivalent to finding the ordered tuple $(d_1, ..., d_k)$ where

$$\begin{aligned}\text{sim}\,(q, d_1) \geq ... \geq \text{sim}\,(q, d_k) \\ \geq \text{sim}\,(q, d) \quad \forall d \notin \{d_1, ..., d_k\}\end{aligned}$$

For dense models, similarity is typically computed as the dot product $\text{sim}\,(q, d) \coloneqq f(q) \cdot f(d)$.

3.2 Referrals

In RAR, we directly use document-to-document relations in the corpus metadata as hard positives, obtaining up to ℓ pairs $(\{q_i(d), d\})_{i=1}^{\ell}$ for each $d \in D$ which are sentences in other documents containing citations or hyperlinks to the current document d.

Aggregating dense representations is usually done via concatenation or taking a sum or average (Izacard and Grave, 2022; Jin et al., 2022; Lin et al., 2022). We experiment with three referral aggregation methods:

1. **Concatenation**: $\tilde{d} \coloneqq [d, q_1(d), ..., q_\ell(d)]$
2. **Mean** $\tilde{f}(d) \coloneqq \frac{1}{\ell+1}\left[f(d) + \sum_i f(q_i(d))\right]$
3. **Shortest path** $\tilde{\text{sim}}\,(q, d) \coloneqq \min\{\text{sim}\,(q, d), (\text{sim}\,(q, q_i(d)))_{i=1}^{\ell}\}$

We discuss and benchmark these methods in Section 4.4 and report main results using the best-performing aggregations.

4 Experiments

4.1 Setup

Tasks We evaluate on paper retrieval (ACL, ArXiv, RefSeer), entity retrieval (DBPedia), open-domain question-answering (NaturalQuestions), and multi-hop open-domain question-answering (HotpotQA). We formulate paper retrieval following past work on local citation recommendation (Gu et al., 2022), with documents as paper titles + abstracts, and queries as citing passages with masked-out citations. The other three tasks follow their setup in the BEIR benchmark (Thakur et al., 2021), with documents as the opening passage of a Wikipedia article and queries as open-ended phrases and questions. Referrals are constructed from citing sentences from the body text of other documents in the corpus, either scientific references or Wikipedia hyperlinks.

Paper retrieval task details For paper retrieval, we partition a corpus of papers into disjoint candidate and evaluation sets — papers in the candidate set represent older, known papers we want to retrieve (for our tasks, candidate papers have publication date ≤ 2018), while papers in the evaluation set represent newer papers whose body text may cite those older papers, each citation inducing a retrieval task with a ground truth (for our tasks, referrals come from papers with publication date ≥ 2019). We compare performance on ACL and

ArXiv papers from the S2ORC corpus (Lo et al., 2020), as well as the open-domain RefSeer corpus, and only include candidate papers that were cited at least once. In-text citations were masked out in both queries and referrals; queries consisted of just the citing sentence, whereas referrals used a 200-token window centered around the masked in-text citation. Documents were augmented with a uniform random sample of up to $\ell = 30$ referrals.

Entity retrieval and QA task details Entity retrieval queries describe named entities, open-domain QA queries give freeform questions with the answer contained in the text of the ground truth document, and multi-hop QA queries give freeform questions whose answers require combined knowledge from multiple documents. We evaluate on the DBPedia, NaturalQuestions, HotpotQA tasks, and compile referrals using the sentences from all pages of other entities that link to the document, parsed from the 2017 English Wikipedia dump via WikiExtractor (Attardi, 2015). We uniformly randomly sample up to $\ell = 30$ referrals per document.

Models For the retriever, we use BM25 (Robertson et al., 2009) as a sparse baseline and (supervised) SimCSE (Gao et al., 2021) and DPR (Karpukhin et al., 2020) as dense baselines. Supervised SimCSE is contrastively fine-tuned from a pretrained BERT on MNLI and SNLI with contradiction pairs as hard negatives without fine-tuning on any retrieval datasets (Gao et al., 2021), and DPR is contrastively fine-tuned on 5 QA datasets (NaturalQuestions, TriviaQA, WebQuestions, CuratedTREC, SQuAD) with BM25 pairs as hard negatives (Karpukhin et al., 2020). We additionally evaluate Specter (Cohan et al., 2020) as an encoder with specialized in-domain training on scientific text: Specter was pretrained on scientific text and then fine-tuned on citing-cited citation pairs from S2ORC (Lo et al., 2020).

4.2 Main results

From Tables 1 and 2, we see that referral augmentation using citing passages from scientific papers and Wikipedia articles **improves sparse and dense retrieval performance on 5 out of 6 tasks**. We get strong gains on paper retrieval tasks **both in an entirely zero-shot setting** (BM25, SimCSE) **as well as for retrievers with in-domain training** (Specter). On both scientific-paper- and Wikipedia-based tasks, we find that referral augmentation yields **more improvement for more structured tasks** for both sparse and dense models. On paper retrieval datasets, the domain-specific ACL dataset on NLP papers saw 90%+ improvements, whereas the ArXiv and RefSeer datasets on papers across scientific fields saw more modest (although still 10%+) ones. Similarly, both BM25 and DPR improve on relatively structured entity retrieval; only BM25 improves on open-domain QA; and neither model improves on multi-hop QA. This degradation of gains from augmentation as tasks become more complex points to a limitation of referral augmentation — its enrichment of document representations comes from short passages from direct neighbors in the referral graph, so it helps more in cases where queries stay semantically close to the direct meaning of the document.

	ACL	ArXiv	RefSeer
BM25	0.265	0.555	0.545
+ RAR	**0.505**	**0.710**	**0.590**
SimCSE	0.160	0.345	0.315
+ RAR	**0.355**	**0.385**	**0.355**
	DBPedia	NaturalQuestions	HotpotQA
BM25	0.40	0.41	0.84
+ RAR	**0.49**	**0.48**	0.84
DPR	0.34	**0.76**	**0.81**
+ RAR	**0.35**	0.73	0.57

Table 1: Retrieval performance (Recall@10) increases across five out of six zero-shot IR tasks with referral augmentation.

	Recall@1	*Recall@10*
Specter	0.084	0.280
+ RAR	**0.106**	**0.341**

Table 2: Referral augmentation additionally helps in-domain retrievers such as Specter. Results shown are on the ACL paper retrieval dataset.

4.3 RAR outperforms other augmentations

In Table 3, we show that **referral augmentation strongly outperforms previous query and document augmentation techniques** exemplified by DocT5Query and Query2Doc. Generative models like DocT5Query fail to capture the more complex text distribution on domains like scientific pa-

	Recall@1	*Recall@10*
BM25	0.13	0.29
+ DocT5Query	0.0	0.155
+ Query2Doc	0.14	0.32
+ RAR	**0.35**	**0.53**

Table 3: Referrals greatly outperform other augmentation techniques. Results shown are on the ACL paper retrieval dataset.

pers and generate qualitatively **nonsensical or trivial queries**, whereas referrals leverage **gold quality reformulations of the paper directly from document-to-document links**.

4.4 Referral aggregation methods

Shortest path aggregation We use a simple, general implementation of shortest path RAR that builds in a black-box way on top of any vector search framework. In the index, we include m copies of each document augmented by a separate referral. Then, for a top-k query, we retrieve the top km queries (or until we have k unique documents); after deduplication, this is equivalent to the top k unique documents under the shortest-path aggregation.

Results We evaluate referral aggregation methods in Table 4 (as defined in Section 3.2). We find that **text concatenation performs the best for BM25** yet poorly for SimCSE, which we hypothesize is due to the repeated text format being out-of-distribution, which affects dense encoders but not inverse term frequencies. **For dense models, mean and shortest path perform the best** for Recall@10 and Recall@1, respectively — we hypothesize this is due to the "smearing" effect of averaging different representations, which leads to more robust document representations at the cost of sacrificing high-precision matches between some referrals and queries. We conclude that for the retrieval task, concatenation for sparse models and mean for dense models results in the best overall performance, and report those for the main results in Table 1.

4.5 Effect of number of referrals

We ablate the number of referrals in paper retrieval, and show in Table 5, that there is a **monotonic improvement in retrieval performance with more referrals**. Note that the improvement has diminishing returns, partially due to a smaller number pool of papers having enough citations to make use of the increased upper bound.

	Recall@1	*Recall@10*
BM25	0.115	0.265
+ RAR $_{\text{concat}}$	**0.200**	**0.505**
+ RAR $_{\text{shortest path}}$	0.093	0.255
SimCSE	0.065	0.160
+ RAR $_{\text{concat}}$	0.060	0.190
+ RAR $_{\text{mean}}$	0.000	**0.355**
+ RAR $_{\text{shortest path}}$	**0.115**	0.265

Table 4: Comparing referral aggregation methods, we find that *concatenation* works best for the sparse model BM25, while *mean* works well for the dense model SimCSE and *shortest-path* achieves the best top-1 performance for SimCSE. Results shown are on the ACL paper retrieval dataset.

	Recall@1	*Recall@10*
BM25	0	0.097
+ RAR ($\leq$ 10 referrals)	0.130	0.371
+ RAR ($\leq$ 20 referrals)	0.156	0.424
+ RAR ($\leq$ 30 referrals)	**0.177**	**0.477**
SimCSE	0.065	0.160
+ RAR ($\leq$ 5 referrals)	0.105	0.295
+ RAR ($\leq$ 30 referrals)	**0.115**	**0.355**

Table 5: Referral augmentation performance increases across the board with the number of referrals used. Results shown are on the ACL paper retrieval dataset.

5 Conclusion

We propose a simple method to leverage document-to-document referrals for retrieval and show strong model- and task-agnostic gains over both base retrievers and other text expansion techniques on a variety of tasks. Going forward, referrals offer a lightweight way to leverage, edit, and aggregate diverse relational perspectives in retrieval corpora.

6 Acknowledgements

We thank Zexuan Zhong and Tianyu Gao for helpful feedback on the paper, as well as anonymous reviewers for their suggestions and comments.

Limitations

The main limitation is that document-to-document links are not always available: referrals can be used with corpora such as academic papers and web-based articles, but not individual passages of books or emails. Here, an effective multi-view retrieval system may need to surface *implicit* referral-like structure, such as the inferred relationships between scenes and characters in a novel, possibly using generative techniques.

In the results, we also discussed limitations of zero-shot RAR augmentation with less structured tasks. To adapt the ideas in RAR to multihop retrieval tasks with more nuanced semantic relationships, it may be useful to synergize ground truth referrals with the in-context generalization capabilities of large language models.

We also note that the concatenation and shortest path aggregation methods lead to longer and more documents, respectively, in linear fashion in ℓ, the number of referrals per augmented document. This is much cheaper than generative expansion methods, and is tractable with $\ell = 30$ and fast max inner product search, but does impose a soft upper bound on the number of referrals it is feasible to take into account, especially for highly cited and linked documents.

Risks

The authors foresee no significant risks with the research presented in this paper.

References

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30:107–117.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers.

Nick Craswell, David Hawking, and Stephen Robertson. 2001. Effective site finding using link anchor information. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 250–257, New York, NY, USA. Association for Computing Machinery.

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*, SIGIR '09, page 227–234, New York, NY, USA. Association for Computing Machinery.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2022. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In *Advances in Information Retrieval*, pages 274–288, Cham. Springer International Publishing.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1265–1268, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard and Edouard Grave. 2022. Distilling knowledge from reader to retriever for question answering.

Di Jin, Rui Wang, Meng Ge, Dongxiao He, Xiang Li, Wei Lin, and Weixiong Zhang. 2022. Raw-gnn: Random walk aggregation based graph neural network.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632.

Marijn Koolen and Jaap Kamps. 2010. The importance of anchor text for ad hoc search revisited. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, page 122–129, New York, NY, USA. Association for Computing Machinery.

Sheng-Chieh Lin, Minghan Li, and Jimmy Lin. 2022. Aggretriever: A simple approach to aggregate textual representation for robust dense passage retrieval.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. S2orc: The semantic scholar open research corpus.

Oliver A. McBryan. 1994. Genvl and wwww: Tools for taming the web. *Computer Networks and Isdn Systems*, 27:308.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th international conference on world wide web*, pages 1291–1299.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Tianyi Zhou, and Daxin Jiang. 2023. Large language models are strong zero-shot retriever.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models.

Thijs Westerveld, Wessel Kraaij, and D. Hiemstra. 2001. Retrieving web pages using content, links, urls and anchors. In *Text Retrieval Conference*.

Jiawen Wu, Xinyu Zhang, Yutao Zhu, Zheng Liu, Zikai Guo, Zhaoye Fei, Ruofei Lai, Yongkang Wu, Zhao Cao, and Zhicheng Dou. 2022. Pre-training for information retrieval: Are hyperlinks fully explored? *arXiv preprint arXiv:2209.06583*.

A Model details

We used pre-trained models from HuggingFace for SimCSE (`sup-simcse-bert-base-uncased`), DPR (`dpr-question_encoder-multiset-base`, `dpr-ctx_encoder-multiset-base`), and Specter (Gao et al., 2021; Karpukhin et al., 2020; Cohan et al., 2020).

B Unifying the RAR formalization with related work

This section unifies the RAR formalization in section 3.1 with related methods in document augmentation and expansion to view methods as different techniques of surfacing positive pairs.

Under the framework defined in section 3.1, the query generation technique DocT5Query (Nogueira et al., 2019) corresponds to generating ℓ hard positive pairs $(\{q_i(d), d\})_{i=1}^{\ell}$ for each $d \in D$, each of which is a question about that document generated by a T5 model (Raffel et al., 2020). For inference, they apply BM25 on the expanded documents $\tilde{d} := [d, q_1(d), ..., q_\ell(d)]$ where $[\cdot, \cdot]$ denotes concatenation.

	Recall@10
BM25 (doc only)	0.643
BM25, doc + anchor texts	0.643
BM25, doc + referrals	0.671
BM25, anchor texts only	0.420
BM25, referrals only	0.614

Table 6: Full hyperlink referrals outperform the ablated anchor text formulation. Results shown are on the ACL paper retrieval dataset.

Similarly, the hypothetical document generation techniques HyDE and Query2Doc (Gao et al., 2022; Wang et al., 2023) correspond to generating ℓ hard positive pairs $(\{q, d_i(q)\}_{i=1}^{\ell}$ *at inference time* for a given query q, each of which is a hypothetical document generated by InstructGPT (Ouyang et al., 2022) to answer the query. For inference, HyDE uses the mean dense encoding between each hypothetical document $\tilde{f}(q) := \frac{1}{\ell+1}[q + \sum_i d_i(q)]$, whereas Query2Doc applies BM25 on the augmented query $\tilde{q} := [q, d_1(q), ..., d_\ell(q)]$ (they use $\ell = 1$, and repeat the original query q a total of $n = 5$ times to emphasize its relative importance).

Under this view, our method differs in that instead of using generative models to produce pairs, we directly leverage ground-truth pair relations in metadata, which often adds high-quality new information.

C Anchor texts vs. referrals

To emphasize that our referral formulation is more effective as well as synergizes better with modern neural retrievers, we ablate the hyperlink referral format for entity retrieval to use just the anchor text, resembling the anchor text setup explored in classical web retrieval (Craswell et al., 2001; Westerveld et al., 2001). In Table 6, we find that augmenting documents with referrals boosts performance, and we can even replace documents entirely with referrals and preserve most of the information value — anchor texts achieve neither.

	Recall@10
SimCSE	0.325
+ RAR (up to 2018)	0.615
+ RAR (up to 2019)	**0.665**

Table 7: Paper retrieval (ACL) on 2020 papers with different referral cutoff years, simulating only having access to referrals up to that year. We find that a larger, updated referral pool improves RAR.

D Referrals allow for training-free modifications to the representation space

One advantage of retriever models over large knowledge-base-like language models is the ability to easily add, remove, and otherwise update documents at inference time with no further fine-tuning. While knowledge editing and patching is an active area of research for large language models (Meng et al., 2023; Cao et al., 2021), all state of the art methods require costly optimization and remain far from matching the convenience and precision of updating a retriever-mediated information store, one reason search engines still dominate the space of internet-scale information organization.

We suggest that referrals naturally extend this property of retrievers, allowing not just documents but *the conceptual relations between documents* and thus the effective representation space to be updated without optimization. On top of adding newly available documents to a retrieval index, we can add their hyperlinks and citations to our collection of referrals, which not only improves retrieval performance on new documents but also *continually improves the representations of older documents* with knowledge of new trends and structure.

To demonstrate the impact of this in a realistic setting, in Table 7 we show the improvement of SimCSE on paper retrieval (evaluating on queries constructed from papers published in 2020) when given additional referrals collected from the metadata of ACL papers released in 2019, compared to only referrals from papers up to 2018. We see that augmenting from an updated pool of referrals improves performance by a significant margin.

Beyond adapting to newly available documents, referrals also open up the possibility of modifying document representations with various down-

stream applications. **This suggests a set of roles for referral modification in retrieval and search systems analogous to how system prompt modification is used with large language models. Human-in-the-loop corrections or additions** can be immediately taken into account by adding them as gold referrals, including adjusting a retrieval system to take trending keywords into account without changing the underlying document content. **Personalized referrals** such as mapping "favorite movie" to "Everything Everywhere All At Once" can also be recorded as a user-specific referral and can be updated at any time. Similarly, **temporary relations** for frequently changing labels such as the “channel of the top trending video on YouTube" or “Prime Minister of the UK" can be kept up to date using referrals. Overall, we are optimistic that referrals unlock new abilities for retrieval systems beyond general improvements to performance.

E Qualitative examples

We include some qualitative examples of paper and entity retrieval and respective retrieved documents for different methods in Table 8.

F Licenses

The ACL and ArXiv queries (in-text citations) and documents (papers) are from S2ORC, which is provided under an ODC-By 1.0 License; RefSeer is provided under a CC BY-NC-SA 3.0 Unported License; and DBPedia is provided under a CC BY-SA 3.0 License. WikiExtractor is available under a GNU Affero General Public License v3.0. All data and artifacts are used as intended.

Query		*[CITATION] showed that BLEU shows high correlation with human scores for grammaticality and meaning preservation and SARI shows high correlation with human scores for simplicity.*		*We leverage the bi-directional Gated Recurrent Units (GRU) [CITATION] to capture the longterm dependency.*
BM25	✗	TerrorCat: a Translation Error Categorization-based MT Quality Metric	✗	Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network
BM25 + RAR	✓	Optimizing Statistical Machine Translation for Text Simplification	✓	Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation
BM25 + DocT5Query	✗	There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction	✗	Deep multi-task learning with low level tasks supervised at lower layers
BM25 + Query2Doc	✗	TerrorCat: a Translation Error Categorization-based MT Quality Metric	✗	Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network

Table 8: Qualitative BM25-based paper retrieval results using different augmentations. In these examples, only RAR retrieval correctly yields the cited paper.