Referral Augmentation for Zero-Shot Information Retrieval

Anonymous ACL submission

Abstract

We propose Referral-Augmented Retrieval (RAR), a simple technique that concatenates document indices with *referrals*, i.e. text from other documents that cite or link to the given document, to provide significant performance gains for zero-shot information retrieval. The key insight behind our method extends an intuition from classical web retrieval: referrals provide a more complete, multi-view representation of a document, much like incoming page links in PageRank provide a comprehensive idea of a webpage's importance. We formulate this classically-rooted intuition as a general augmentation and find that it empirically works across various new domains and retrieval methods, outperforming modern generative text expansion techniques such as DocT5Query (Nogueira et al., 2019) and Query2Doc (Wang et al., 2023) — a 37% and 21% absolute improvement on ACL paper retrieval Recall@10, respectively, while also eliminating expensive model training and inference. We also analyze different methods for multi-referral aggregation and show that RAR enables up-to-date information retrieval without re-training. We believe RAR can help revive and re-contextualize this classic information retrieval intuition in the age of neural retrieval, unlocking new retrieval gains by combining untapped corpus structure with the semantic advantages of modern pretrained transformers.

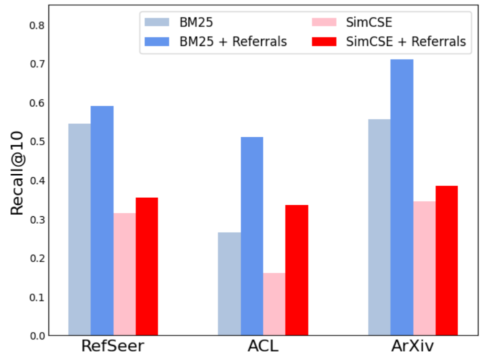

Figure 1: Our referral augmentation method improves zero-shot document retrieval across a variety of models and datasets.

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.

While the zero-shot setting lacks query-document pairs, our key insight is to leverage intra-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 intra-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

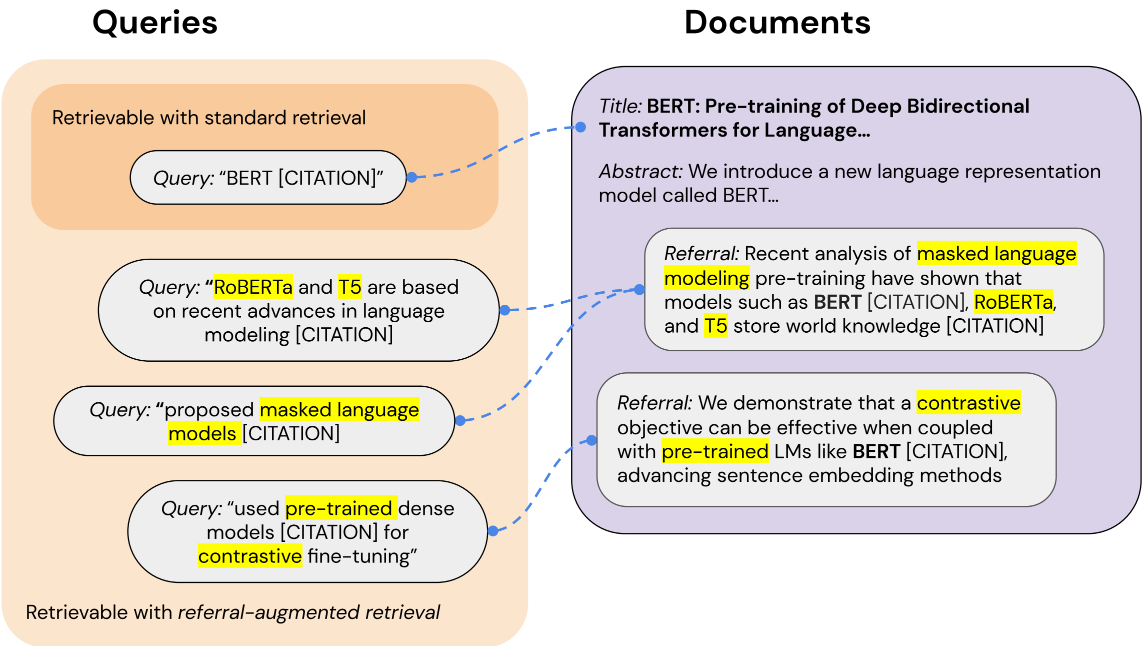

Figure 2: Illustration of the Referral-Augmented Retrieval (RAR) process. RAR augments text from documents that refer to the original document into its index (right), which allows it to correctly retrieve the target document for a wider range of queries (left) compared to standard methods. This example uses text around citations as queries, from the citation recommendation task (Gu et al., 2022).

with referral links like the Semantic Scholar citation graph (Lo et al., 2020) and Wikipedia entity graph (Hasibi et al., 2017).

For both the paper retrieval and entity retrieval settings, we find that RAR significantly improves zero-shot retrieval performance for both sparse and dense models. For instance, RAR outperforms generative text expansion techniques such as DocT5Query (Nogueira et al., 2019) and Query2Doc (Wang et al., 2023) by up to 37% and 21% Recall@10, respectively, on ACL paper retrieval from the S2ORC corpus (Lo et al., 2022). Moreover, RAR's augmentation occurs entirely at indexing time and hence allows for a training-free method to update a retrieval system with new views of existing documents (e.g., a trending news story that causes users to search for a public figure by the name of the scandal they were in), recontextualizing the strengths of this classical idea in new ways (more in Section 5.2). We also find that our method scales well as the number of referrals increases and is easy to update.

Another example of insights from recontextualization comes from comparing RAR to popular modern query and document expansion techniques (Nogueira et al., 2019; Gao et al., 2022; Wang et al., 2023). Text expansion techniques effectively surface *hard positives*, passages that are very lexically different but semantically equivalent, including conceptual transformations (e.g., mapping a claim to a piece of contradictory evidence), the addition of new information, and alternative formulations with different word choice or scope. While some of these transformations are theoretically learnable, existing dense retrievers are often not robust to them, so explicitly augmenting documents and queries with their equivalent counterparts significantly improves the encoded representations. As an added bonus, the text-to-text nature of these hard positive pairs allows them to be both model-agnostic and interpretable. This observation motivates further research into improving retrieval not by training a more expressing encoder, but *by simply discovering more hard positives.*

2 Related Work

Sparse and dense retrieval Following the success of BERT (Devlin et al., 2019), a variety

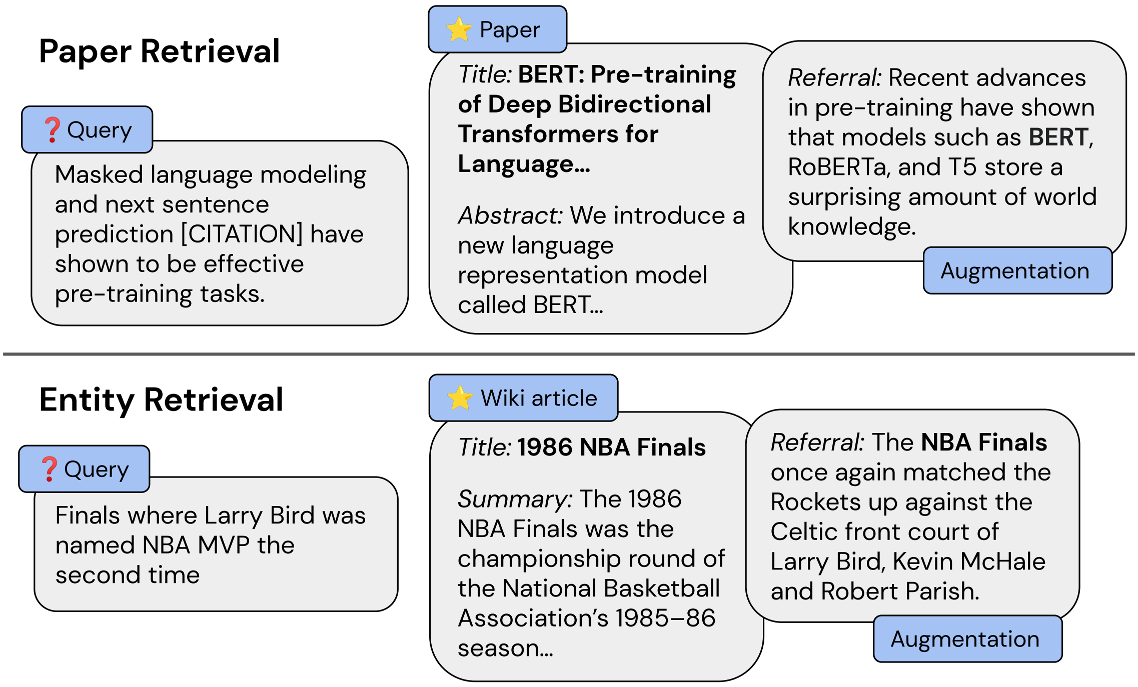

Figure 3: We evaluate referral augmentation on zero-shot paper retrieval, retrieving papers given masked in-text citations, (top) and entity retrieval, retrieving wiki articles on each titular entity given free text queries about the entity (bottom).

of BERT-based dense encoder models have been proposed for information retrieval. Karpukhin et al. (2020) propose DPR, fine-tuning on query-document pairs from MS MARCO (Bajaj et al., 2018); Gao et al. (2021) propose SimCSE, fine-tuning using supervision from NLI datasets with entailment pairs as positives and contradiction pairs as hard negatives; and Izacard et al. (2021) propose Contriever, fine-tuning using random crops and MoCo (He et al., 2020) to scale to a large number of negatives. However, Thakur et al. (2021) show that term-frequency sparse methods like BM25 remain a strong baseline in the zero-shot IR setting.

Hyperlinks 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 all incoming anchor texts for a page into an "anchor document" (Craswell et al., 2001), and this method is refined for web search tasks in the following years using ad hoc combinations of anchor and content-based rankings as well as multiple retrieval passes for query expansion (Westerveld et al., 2001; Eiron and McCurley, 2003; Arguello et al., 2021; Koolen and Kamps, 2010; Dou et al., 2009). Twenty years after these seminal works, we find that longer passage-length referrals improve the context of deep pretrained transformer encoders in analogous ways to the gains of statistical rankers from word- and phrase-length anchor texts (Craswell et al., 2001; Westerveld et al., 2001). Compared to 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. (we empirically compare anchor texts and full referrals in Section B in the Appendix) 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 One previous line of work explores using hyperlinks and citations for *training* retrievers, using referrals indirectly as a way of constructing a dataset of paired passages for contrastive learning. Entity retrieval models Mitra et al. (2017) and Wu et al. (2022) explore pre-training using the anchor text portion of a linking sentence as a pseudo-query for query-document pre-training, among other pre-training objectives, and explore different kinds of relevance classes based on whether the link is mutual. State-of-the-art paper retrieval approaches (Gu et al., 2022) (Cohan et al., 2020) similarly fine-tune using (citing paper's title + abstract + citing passage, cited paper's title + abstract) pairs. In contrast, we focus on using hyperlinks and citations to build *training-free* document augmentations that work with any off-the-shelf encoder. This direction is also orthogonal to our work, since we find empirically that a stronger embedding space (e.g. trained via data mined from anchor text) can still benefit from our RAR method of document expansion, as seen in Table 6 in the Appendix.

Query and document expansion Query expansion techniques were originally proposed to decrease the lexical gap between queries and documents, using relevance feedback as well as external knowledge banks like WordNet (Miller, 1995), whereas document expansion techniques such as Doc2Query and DocT5Query (Nogueira et al., 2019) were intended to add additional context and surface key terms. Some work also explores sparse retrievers with learned document term weights (Formal et al., 2021) and late interaction models (Khattab and Zaharia, 2020), which can be seen as performing implicit document expansion. However, most state-of-the-art dense retrievers (Gao et al., 2021; Karpukhin et al., 2020) do not perform any expansion, and in this work we have shown that they benefit significantly from referrals.

Model updating and editing An ongoing line of work (Meng et al., 2023; Cao et al., 2021) studies fact editing for language models, which are resource-intensive to modify and trained on data that quickly becomes outdated. Retrieval systems trivially admit document edits and the addition of new documents without training, and we have found that hard negatives and referrals extend this property to support multiple document views. These benefits can reach end-to-end generation via retriever-augmented language models (Ram et al., 2023; Guu et al., 2020).

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 a 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 between the encodings of queries, where the encoder is shared:

$$\text{sim}(q, d) := f(q) \cdot f(d)$$

We can formally define a hard positive as a pair of highly relevant passages $\{x_1, x_2\}$ that should be mapped to the same point in embedding space, which in effect imposes a correction on top of a given encoder f where $f(x_1) \neq f(x_2)$. We discuss a unifying viewpoint on other expansion methods (Doc2Query, HyDE, Query2Doc (Nogueira et al., 2019; Gao et al., 2022; Wang et al., 2023)) in Section 6 in the Appendix.

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. We experiment with three different referral integration methods:

1. **Concatenation**: $\tilde{d} := [d, q_1(d), ..., q_\ell(d)]$
2. **Mean** $\tilde{f}(d) := \frac{1}{\ell+1}\left[f(d) + \sum_i f(q_i(d))\right]$

3. **Shortest path** $\tilde{\text{sim}}(q, d) := \min\{\text{sim}(q, d), (\text{sim}(q, q_i(d)))_{i=1}^{\ell}\}$

We find in Section 5.2 that for sparse models, concatenation performs the best, while for dense models, mean aggregation performs the best, although shortest path achieves the best top 1 accuracy (Recall@1) since it preserves the high granularity of separate referrals, and use these settings when reporting overall results.

4 Experiments

4.1 Setup

Paper retrieval Paper retrieval is the task of retrieving papers most likely to be cited in a given passage. 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, 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. Following the classic setup of *local citation recommendation* (LCR) (Gu et al., 2022), we represent each candidate paper via its concatenated title and abstract, and construct a query from each sentence in an evaluation papers referencing a candidate paper (with the citation masked). To evaluate the effects of augmenting a candidate document at indexing time, we compile referrals consisting of citing sentences in *other candidate papers*.

We compare performance with and without augmentation on ACL and ArXiv papers from the S2ORC corpus (Lo et al., 2020), as well as the open-domain RefSeer corpus. ACL and ArXiv paper retrieval tasks were partitioned such that papers published in 2018 or before comprised the candidate set, and papers in 2019 comprised the evaluation set, filtering to 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 Entity retrieval is the task of retrieving the most relevant entities from a knowledge base given a text query. We evaluate on the DBPedia entity retrieval task, which represents each entity (associated with a Wikipedia page) via its concatenated name and summary, and contains freeform text queries. To augment a candidate document, we compile referrals consisting of sentences from the pages of other entities that link to the the document. We used the 2017 English Wikipedia dump preprocessed with WikiExtractor (Attardi, 2015) and extract hyperlinks via a HTML parser, again including a random sample of up to 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), contrastively fine-tuned BERT encoders, as dense baselines. Supervised SimCSE is contrastively fine-tuned from a pre-trained BERT on MNLI and SNLI with contradiction pairs as hard negatives (Gao et al., 2021), and DPR is contrastively fine-tuned on 5 QA datasets (NQ, TriviaQA, WebQuestions, CuratedTREC, SQuAD) with mined BM25 pairs as hard negatives (Karpukhin et al., 2020). We also evaluate on BM25 + CE, which adds a cross-encoder to the BM25 model (Wang et al., 2020) and was found to be the best-performing zero-shot retriever from the BEIR evaluation (Thakur et al., 2021). For paper retrieval, we also evaluate the effect of using referrals with Specter (Cohan et al., 2020), a domain-specific encoder pre-trained and fine-tuned on scientific text.

4.2 Results

Paper retrieval From Table 1, we see that a retriever augmented with referrals outperforms the base retriever for all sparse and dense models, with significant improvement on both Recall@1 and Recall@10 on all datasets (including an extremely large 100% improvement on ACL) for BM25 + RAR compared to regular BM25. We see that alongside surfacing more relevant information to increase recall, referrals also greatly increase the specificity to generate much better top-1 retrieved candidates, pointing to the fact that referring citations referencing a paper are often more clear, concise, and well-specified than the abstract of the paper itself.

	RefSeer	ACL	ArXiv
		Recall@10 (Recall@1)	
BM25	0.545 (0.260)	0.265 (0.115)	0.555 (0.335)
+ RAR	**0.590** (**0.335**)	**0.505** (**0.200**)	**0.710** (**0.430**)
SimCSE	0.315 (0.095)	0.160 (0.065)	0.345 (**0.140**)
+ RAR	**0.355** (**0.155**)	**0.355** (**0.115**)	**0.385** (0.120)

Table 1: Paper retrieval results with citation referrals. RAR greatly improves paper retrieval performance for both sparse and dense models on all metrics, sometimes doubling the absolute performance.

	nDCG@1	*nDCG@10*	*Recall@10*
BM25	0.4030	0.2739	**0.1455**
+ RAR	**0.4851**	**0.2799**	0.1348
BM25 + CE	0.4254	0.3282	0.1798
+ RAR	**0.4478**	**0.3283**	**0.1949**
DPR	0.3350	0.2559	0.1562
+ RAR	**0.3538**	**0.2610**	**0.1612**

Table 2: Entity retrieval results with hyperlink referrals, on the DBPedia task. RAR improves entity retrieval performance on both sparse and dense models.

Entity retrieval We evaluate model performance with and without referrals in Table 2. We see that referrals again significantly elevate performance for both sparse and dense models across the board. The gain is particularly large for nDCG@1, which we hypothesize is due to the occasionally extremely high similarity of referring sentences with some queries.

We note that hyperlink referrals do not increase performance as much as the respective citation referrals on the paper retrieval task, suggesting that linking sentences may be less consistent and less directly informative than citing ones. Intuitively, different citations of a given scientific work are typically similar in spirit, while the relevance relations implied by different hyperlinks may be more tangential. However, this is not necessarily a fair comparison, as the Wikipedia-based query and corpus distributions also vary much more and encompass more diverse fields of knowledge.

5 Analysis

5.1 Referrals outperform other augmentations

In Table 3, we show that referral augmentation strongly outperforms 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 papers and generate qualitatively nonsensical or trivial queries, whereas referrals leverage gold quality reformulations of the paper directly from document-to-document links.

5.2 Referral aggregation methods

Aggregating dense representations is a well-known problem (Izacard and Grave, 2022; Jin et al., 2022; Lin et al., 2022), and is usually resolved via concatenation or taking a sum or average. We propose three such methods: text concatenation, mean representation, and shortest path (details in section 3.2), which we will denote by $\text{referrals}_{\text{concat}}$, $\text{referrals}_{\text{mean}}$, $\text{referrals}_{\text{sp}}$. Note that BM25 does not support mean aggregation since it does not yield vector embeddings.

We include the shortest path method as a means

	Recall@1	*MRR@10*	*Recall@10*
BM25	0.13	0.177	0.29
+ RAR	**0.35**	**0.4088**	**0.53**
+ DocT5Query	0.0	0.036	0.155
+ DocT5Query + RAR	0.345	0.4022	0.525
+ Query2Doc	0.14	0.1940	0.32
+ Query2Doc + RAR	**0.38**	**0.4279**	0.52

Table 3: Paper retrieval, referrals vs. other augmentation techniques (Recall@10). We bold the best result on any single augmentation strategy, as well as any results on stacked augmentations that show further gains over that single augmentation. Overall, we find that referrals greatly outperform other augmentation techniques, and further that referrals can stack with Query2Doc to achieve even better performance.

to take advantage of different referrals representing distinct views of a given document that should not necessarily be aggregated as a single mean embedding — while citations are fairly consistent, hyperlinks to a given article sometimes focus on unrelated aspects of its content (e.g. referencing a famous painting by its painter vs. by its host museum) which may be best represented by different locations in query space.

Results We evaluate them in Table 4 and find that text concatenation performs the best for BM25 but poorly for SimCSE, which we hypothesize is due to the fact that repetition and concatenation of text improves the approximation of a target query (inverse term frequency) distribution for BM25, but results in a distorted dense representation since dense models approach text sequentially and in particular a long string of referring sentences in a row is very much out of their training distribution.

For dense models, mean and shortest path aggregation performs the best for Recall@10 and Recall@1, respectively. We hypothesize that this is due to the "smearing" effect of averaging many different representations which leads to more robust document representations generally, but possibly at the cost of the high precision resulting from some referrals being an almost-perfect match for some queries at evaluation time. We conclude that for the retrieval task, concatenation for sparse models and mean for dense models results in the best overall performance, and use this configuration when reporting the main results in Table 1.

5.3 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 5 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.[1] We see

[1] Specifically, we add the in-text citations of later layers to the pool of referrals, from which we randomly resample up to

	Recall@1	*MRR@10*	*Recall@10*
BM25	0.115	0.157	0.265
+ RAR $_{concat}$	**0.200**	**0.2677**	**0.505**
+ RAR $_{sp}$	0.093	0.1406	0.255
SimCSE	0.065	0.0869	0.160
+ RAR $_{concat}$	0.060	0.0989	0.190
+ RAR $_{mean}$	0.000	0.111	**0.355**
+ RAR $_{sp}$	**0.115**	**0.158**	0.265

Table 4: Paper retrieval results, comparing different 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.

	ACL
SimCSE	0.325
+ RAR (up to 2018)	0.615
+ RAR (up to 2019)	**0.665**

Table 5: Paper retrieval on 2020 papers with different referral cutoff years (Recall@10). We find that an updated referral pool improves referral-augmented retrieval.

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 relationships for a variety of applications. **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. Clearly, we find that referrals unlock new abilities for retrieval systems beyond general improvements to performance.

$\ell = 30$ per document when building the retrieval index; the total number of citations is unchanged for most documents that already have 30 referrals available from the original dataset.

6 Conclusion

We propose a simple method to capture implicit hard positives using intra-document citations and hyperlinks as *referrals* to provide alternate views of a given document, and show that referral augmentation yields strong model- and task-agnostic gains for zero-shot retrieval that outperforms previous text expansion techniques while also being less expensive. We also explore applications of hard positives as training-free modifications to the representation space, allowing new views of documents to be dynamically added to reflect updated world context, human-in-the-loop corrections, and personalized and temporary labels for documents.

One perspective on our referral augmentation results is evidence that an index that incorporates multiple views per document may be better suited for the retrieval of high-quality, atomic documents that may nevertheless each be relevant to a variety of different situations. It is also apparent that often these views may not be apparent from the document text itself — for example, a paper may be commonly referenced as the progenitor of a follow-up work, of which it obviously has no knowledge. Our work offers a preliminary look at a simple way to collect some of these nonobvious multiple views from the corpus itself, as well as the aggregation problem that subsequently arises. Our work thus suggests that the more general problem of fully capturing these distinct facets of each document — and efficiently determining which facet is most relevant to a given query — may be an important next step for robust retrieval.

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.

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. Thus, the augmentation trades off memory and speed for more relevant retrieved documents. This is tractable (and insignificant compared to the costs of generative expansion methods) with our choice of $\ell = 30$ and fast max inner product search algorithms, 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

Jaime Arguello, Jonathan L. Elsas, Jamie Callan, and Jaime Carbonell. 2021. Document representation and query expansion models for blog recommendation. *Proceedings of the International AAAI Conference on Web and Social Media*, 2(1):10–18.

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

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. 2018. Ms marco: A human generated machine reading comprehension dataset.

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.

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

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.

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*, SIGIR '03, page 459–460, New York, NY, USA. Association for Computing Machinery.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning.

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

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.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert.

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.

Chun Hei Lo, Wai Lam, and Hong Cheng. 2022. Semantic composition with PSHRG for derivation tree reconstruction from graph-based meaning representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5425–5439, Dublin, Ireland. Association for Computational Linguistics.

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.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models.

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

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.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

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 Unifying perspective on expansion methods

Under the framework defined in section 3, 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.

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).

B Referral augmentation for task-specific models

We additionally compare against Specter, a state-of-the-art task-specific paper retrieval model with pretraining on scientific text and contrastive fine-tuning specifically on pairs of papers that cite each other (Cohan et al., 2020). We find in Table 6 that referral augmentation still helps by a large margin for the task-specific model, so we consider the uses of citations for referral augmentation and training orthogonal.

	Recall@1	*MRR@10*	*Recall@10*
Specter	0.084	0.136	0.280
+ RAR	**0.106**	**0.169**	**0.341**

Table 6: Paper retrieval results for Specter on ACL. We find that referral augmentation helps even when referrals were used for task-specific model training.

	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 7: Full hyperlink referrals outperform the ablated anchor text formulation.

C Anchor texts vs. referrals

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 7, 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.

D Effect of number of referrals

We ablate the number of referrals in paper retrieval, and show in Table 9, 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 actually having enough citations to benefit.

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.

	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 9: Paper retrieval results on different numbers of referrals on ACL. We find that performance increases across the board with the number of referrals used.

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.