Generative Retrieval Meets Multi-Graded Relevance

Yubao Tang[1], Ruqing Zhang[1], Jiafeng Guo[1], Maarten de Rijke[2], Wei Chen, Xueqi Cheng[1]

[1]CAS Key Lab of Network Data Science and Technology, ICT, CAS; University of Chinese Academy of Sciences

[2]University of Amsterdam

{tangyubao21b,zhangruqing,guojiafeng, chenwei2022,cxq}@ict.ac.cn

m.derijke@uva.nl

Abstract

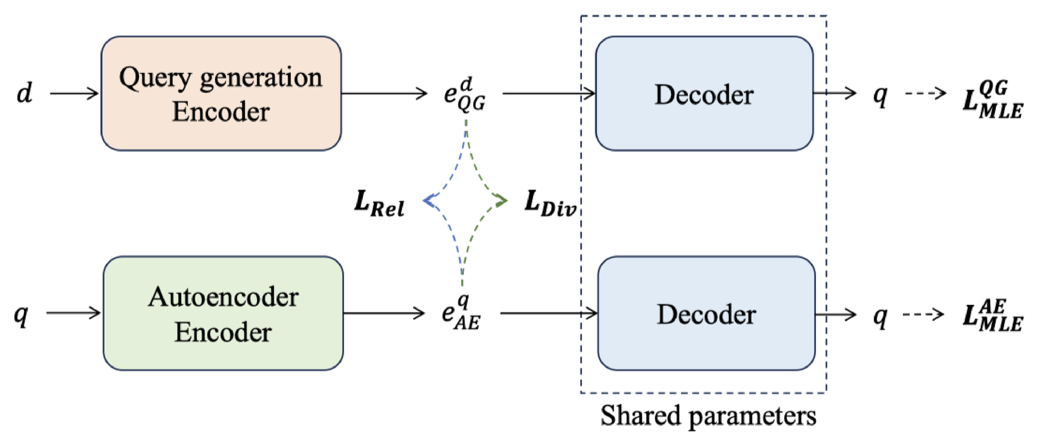
- Generative retrieval (GR) represents a novel approach to information retrieval. It uses an encoder-decoder architecture to directly produce relevant document identifiers (docids) for queries. While this method offers benefits, current approaches are limited to scenarios with binary relevance data, overlooking the potential for documents to have multi-graded relevance. Extending GR to accommodate multi-graded relevance poses challenges, including the need to reconcile likelihood probabilities for docid pairs and the possibility of multiple relevant documents sharing the same identifier.
- To address these challenges, we introduce a framework called GRaded Generative Retrieval (GR^2). GR^2 focuses on two key components: ensuring relevant and distinct identifiers, and implementing multi-graded constrained contrastive training. First, we create identifiers that are both semantically relevant and sufficiently distinct to represent individual documents effectively. Second, we incorporate information about the relationship between relevance grades to guide the training process. Extensive experiments on datasets with both multi-graded and binary relevance demonstrate the effectiveness of GR^2.

Introduction

- **Generative retrieval** (GR) encodes all information in a corpus into the model parameters, and produces a ranked list based on a single parametric model
- Current work on GR mainly focuses on binary relevance scenarios, where a binary division into relevant and irrelevant categories is assumed [20, 52, 83]
- The standard Seq2Seq objective, via maximizing likelihood estimation (MLE) of the output sequence with teacher forcing, has been used extensively in GR due to its simplicity
- In real-world search scenarios, documents may have different degrees of relevance [18, 71, 72, 82]
- A straightforward approach to extending GR to multiple grades, involves having the GR model generate the likelihood of docids with higher relevance grades being greater than that of lower grades
- However, the variation in docid lengths may lead to smaller likelihood scores for longer docids
- Besides, essential topics in multi-graded relevant documents may be similar, emphasizing the need for a one-to-one correspondence between document content and its identifier to ensure distinctness
- Consequently, harnessing a GR model's capabilities for multi-graded relevance ranking in a relatively succinct manner remains an non-trivial challenge
- **Contribution:** A novel GRaded Generative Retrieval (GR^2) framework
  - Docid: Enhance docid distinctness while ensure its relevance to document semantics
  - Relevance: A multi-graded constrained contrastive loss to capture different relevance grades
  - Application: Two learning scenarios, i.e., supervised learning and pre-training

Approach - Docid

- **Docid design: regularized fusion approach**
  - **Key idea: (i) use** pseudo-queries generated from the document as docids with a query generation (QG) model; (ii) : jointly optimize the relevance and distinctness that fuses the latent space of a QG model, i.e., a docid generation model, and that of an autoencoder (AE) model



  - **Relevance regularization term**: (i) encourage the representation of a document and that of the corresponding docid (i.e., pseudo-query) to be close to each other in the shared latent space. (ii) increase the distance between the representation of a document and that of irrelevant docids associated with other documents

$$\mathcal{L}_{Rel}(Q, D_Q; \theta_{QG}, \theta_{AE}) = -\frac{1}{|Q|}\sum_{q\in Q, d\in D_Q} \frac{\exp(sim(e^d_{QG}, e^q_{AE}))}{\exp(sim(e^d_{QG}, e^q_{AE})) + \zeta},$$

where $\zeta = \sum_{d\in D_Q, \bar{q}\in Q, \bar{q}\neq q} \exp(sim(e^d_{QG}, e^{\bar{q}}_{AE}))$

  - **Distinctness regularization term:** To enhance the distinctness between documents and between docids, we push away the representations of different documents in the document space and, simultaneously, push away the representations of different docids in the docid space

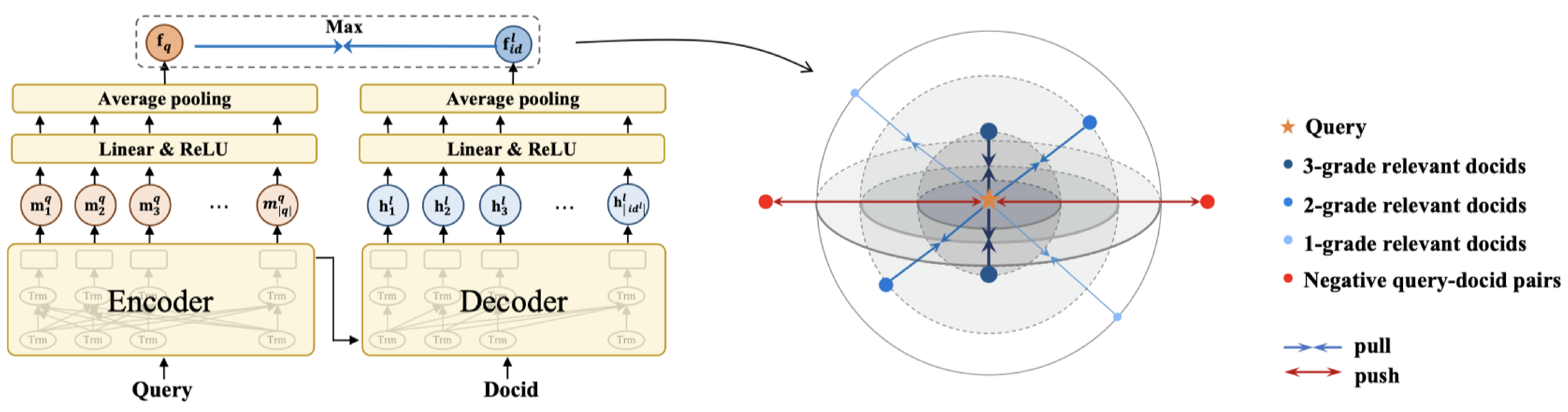
$$\mathcal{L}_{Div}(\cdot) = \sum_{d,\bar{d}\in D_Q, d\neq \bar{d}} \frac{sim(e^d_{QG}, e^{\bar{d}}_{QG})}{|Q|(|Q|-1)} + \sum_{q,\bar{q}\in Q, q\neq\bar{q}} \frac{sim(e^q_{AE}, e^{\bar{q}}_{AE})}{|Q|(|Q|-1)} - \sum_{q\in Q, d\in D_Q} \frac{sim(e^d_{QG}, e^q_{AE})}{|Q|}$$

Approach – MGCC loss

- **Multi-graded constrained contrastive loss**: force positive pairs closer together in the representation space, but the magnitude of the force is dependent on the relevance grade
- **Grade penalty:** To distinguish between multiple positive pairs, our key idea is to apply higher penalties to positive pairs constructed from higher grades, forcing them closer than negative pairs constructed from lower grades

$$\mathcal{L}_{Pair}(q, id^l;\theta) = \log\frac{\exp(sim(\mathbf{f}_q, \mathbf{f}^l_{id})/\tau)}{\sum_{a\in A_q}\mathbb{1}_{[\mathbf{f}_q\neq\mathbf{f}_a]}\exp(sim(\mathbf{f}_q, \mathbf{f}_a)/\tau)} \qquad \sum_{q\in Q}\frac{1}{L}\sum_{l=1}^{L}\frac{-\lambda_l}{|I_{D^l_q}|}\sum_{id^l\in I_{D^l_q}}\mathcal{L}_{Pair}(q, id^l;\theta),$$

- **Grade constraint:** A class higher in the hierarchy cannot have a lower confidence score than a class lower in the ancestry sequence: $\mathcal{L}_{Max}(l, q, id^l) = \max_{(q, id^l;\theta)} \mathcal{L}_{Pair}(q, id^l;\theta)$
- **MGCC loss:** $\mathcal{L}_{MGCC}(\cdot) = \sum_{q\in Q}\frac{1}{L}\sum_{l=1}^{L}\frac{-\lambda_l}{|I_{D^l_q}|}\sum_{id^l\in I_{D^l_q}} \max(\mathcal{L}_{Pair}(q, id^l;\theta), \mathcal{L}_{Max}(l+1, q, id^{l+1};\theta))$



Approach - Optimization

- **Supervised learning** (GR^{2S}) $\mathcal{L}_{total}(Q, D, I_D;\theta) = \gamma\mathcal{L}_{MGCC}(Q, I_{D_Q};\theta) + \mathcal{L}^q_{MLE}(Q, I_{D_Q};\theta) + \mathcal{L}^d_{MLE}(D, I_D;\theta),$
- **Pre-training and fine-tuning** (GR^{2P})**:** (i) To construct pre-training data, we use the English Wikipedia [87] to build a set of pseudo-pairs of queries and docids. (ii)We use the unique titles of Wikipedia articles as the docids for pre-training and assume that a random sentence in the abstract can be viewed as a representative query of the article
  - grade 4: the Wikipedia article from which the query is sampled, is regarded as the most relevant document
  - grade 3: We use the See Also section of a Wikipedia article in which hyperlinks link to other articles with similar or comparable information, which is mainly written manually. If there exists no See Also section, we use a similar section, i.e., the Reference section
  - grade 2 and grade 1: Besides the See Also section, some hyperlinks link to pages that describe the concept of some entities in detail. We randomly sample several anchor texts from the first section and other sections, respectively, and regard the linked target pages as grade 2 and grade 1 relevant documents, respectively

Experimental settings & results

Table 1: Experimental results on datasets with multi-graded relevance. Results denoted with ⋆ are from [34, 55]. And ∗,†, ‡ and ≀ indicate statistically significant improvements over the best performing SR baseline QLM, the DR baseline PseudoQ, the GR baseline RIPOR, and all the baselines, respectively ($p \leq 0.05$).

Methods	Gov 500K				ClueWeb 500K				Robust04			
	nDCG		P	ERR	nDCG		P	ERR	nDCG		P	ERR
	@5	@20	@20	@20	@5	@20	@20	@20	@5	@20	@20	@20
BM25	0.4984	0.4819	0.5374	0.1848	0.2579	0.2417	0.3471	0.1362	-	0.4193⋆	**0.3657**⋆	0.1140⋆
DocT5query	0.3936	0.3861	0.4177	0.1258	0.2071	0.1631	0.2604	0.0821	0.3613	0.3229	0.3023	0.1075
QLM	0.4987	0.4822	0.5379	0.1851	0.2582	0.2423	0.3475	0.1365	0.4121	0.4195	0.3658	0.1143
SPLADE	0.4370	0.4146	0.4445	0.1575	0.2272	0.2155	0.3050	0.1109	0.4031	0.3640	0.3192	0.1088
RepBERT	0.3101	0.3351	0.4305	0.1446	0.2624	0.2431	0.3650	0.1663	0.2725	0.2212	0.1686	0.0812
DPR	0.3236	0.3408	0.4417	0.1597	0.2614	0.2576	0.3754	0.1737	0.2873	0.2316	0.1788	0.0873
PseudoQ	0.4168	0.4383	0.5134	0.1801	0.2752	0.2704	0.3926	0.1815	0.4072	0.3577	0.2823	0.0927
ANCE	0.4152	0.4379	0.5129	0.1794	0.2743	0.2696	**0.3919**	0.1809	0.4069	0.3573	0.2820	0.0921
DSI-Num	0.2484	0.2647	0.3237	0.1052	0.1942	0.1690	0.2520	0.1063	0.2699	0.2028	0.1524	0.0711
DSI-Sem	0.2497	0.2745	0.3392	0.1215	0.2004	0.1977	0.2669	0.1143	0.2711	0.2135	0.1649	0.0737
SEAL	0.3914	0.3255	0.4418	0.1592	0.2683	0.2293	0.2927	0.1305	0.2823	0.2287	0.1654	0.0855
DSI-QG	0.4566	0.4365	0.4602	0.1702	0.2722	0.2556	0.3625	0.1783	0.4089	0.3703	0.3267	0.1032
NCI	0.4635	0.4473	0.4722	0.1882	0.2783	0.2631	0.3734	0.1896	0.4096	0.3786	0.3349	0.1052
Ultron-PQ	0.4658	0.4496	0.4775	0.1911	0.2798	0.2652	0.3758	0.1904	0.4103	0.3797	0.3352	0.1063
LTRGR	0.4663	0.4517	0.4783	0.1923	0.2805	0.2664	0.3762	0.1916	0.4109	0.3805	0.3358	0.1071
GenRRL	0.4669	0.4524	0.4789	0.1928	0.2812	0.2669	0.3768	0.1921	0.4112	0.3810	0.3362	0.1078
GenRet	0.4672	0.4528	0.4792	0.1931	0.2824	0.2671	0.3770	0.1925	0.4116	0.3812	0.3365	0.1081
NOVO	0.4675	0.4531	0.4796	0.1935	0.2827	0.2674	0.3372	0.1928	0.4119	0.3816	0.3369	0.1084
RIPOR	0.4713	0.4578	0.4831	0.1978	0.2835	0.2707	0.3401	0.1963	0.4142	0.3849	0.3404	0.1093
GR^{2S}	0.4869≀	0.4784≀	0.5364≀	0.2125≀	0.2886∗†	0.2791≀	0.3788∗‡	0.2016∗†	0.4197∗†	0.3983≀	0.3471≀	0.1097†
GR^{2P}	**0.5095**≀	**0.4912**≀	**0.5506**≀	**0.2167**≀	**0.3034**≀	**0.2969**≀	0.3871∗‡	**0.2026**≀	**0.4301**≀	**0.4205**≀	0.3568≀	**0.1196**≀

Table 2: Experimental results on datasets with binary relevance. And ∗,†, ‡ and ≀ indicate statistically significant improvements over the best performing SR baseline DocT5query or SPLADE, the DR baseline PseudoQ, the GR baseline RIPOR and all the baselines, respectively ($p \leq 0.05$).

Methods	MS 500K				NQ 320K			
	MRR		Hits		MRR		Hits	
	@3	@20	@1	@10	@3	@20	@1	@10
BM25	0.2171	0.2532	0.2385	0.3969	0.1456	0.1875	0.2927	0.6016
DocT5query	0.3378	0.3561	0.3489	0.5773	0.2612	0.2859	0.3913	0.697
QLM	0.2746	0.2805	0.2852	0.4593	0.2625	0.2864	0.3927	0.6979
SPLADE	0.3246	0.3483	0.3353	0.5637	0.3057	0.3404	0.4253	0.7146
RepBERT	0.3029	0.3382	0.3287	0.5233	0.3135	0.3421	0.4542	0.7275
DPR	0.3095	0.3264	0.3215	0.5432	0.3172	0.3493	0.5020	0.7812
PseudoQ	0.3342	0.3528	0.3452	0.5736	0.3253	0.3582	0.5271	0.7952
ANCE	0.3330	0.3520	0.3446	0.5729	0.3215	0.3576	0.5263	0.7931
DSI-Num	0.2159	0.2798	0.2676	0.4440	0.2286	0.2793	0.2185	0.4571
DSI-Sem	0.2229	0.2847	0.2753	0.4832	0.2581	0.3084	0.2740	0.5660
GENRE	-	-	-	-	0.3268	0.3467	0.2630	0.7120
SEAL	0.2977	0.3110	0.3072	0.5163	0.3367	0.3658	0.2630	0.7450
DSI-QG	0.3271	0.3457	0.3352	0.5749	0.3613	0.3868	0.6349	0.8236
NCI	0.3317	0.3566	0.3365	0.5833	0.3657	0.4053	0.6424	0.8311
Ultron-PQ	0.3326	0.3575	0.3379	0.5851	0.3663	0.4059	0.6461	0.8345
LTRGR	0.3354	0.3583	0.3381	0.5859	0.3692	0.4078	0.6511	0.8489
GenRRL	0.3359	0.3587	0.3389	0.5863	0.3698	0.4086	0.6528	0.8533
GenRet	0.3362	0.3591	0.3393	0.5867	0.3702	0.4095	0.6542	0.8567
NOVO	0.3371	0.3602	0.3405	0.5869	0.3724	0.4136	0.6613	0.8624
RIPOR	0.3384	0.3626	0.3421	0.5873	0.3741	0.4173	0.6638	0.8667
GR^{2S}	0.3489≀	0.3714≀	0.3515†‡	0.6126≀	0.3813≀	0.4299≀	0.6724≀	0.8713∗†
GR^{2P}	**0.3597**≀	**0.3835**≀	**0.3821**≀	**0.6405**≀	**0.3937**≀	**0.4418**≀	**0.6832**≀	**0.8825**≀

Table 4: Comparison between GR methods and the full-ranking baseline

Methods	Gov 500K	MS 500K
	nDCG@5	MRR@20
RIPOR	0.4713	0.3626
GR^{2P}	0.5095	0.3835
BM25+monoBERT	**0.6953**∗	**0.5862**∗

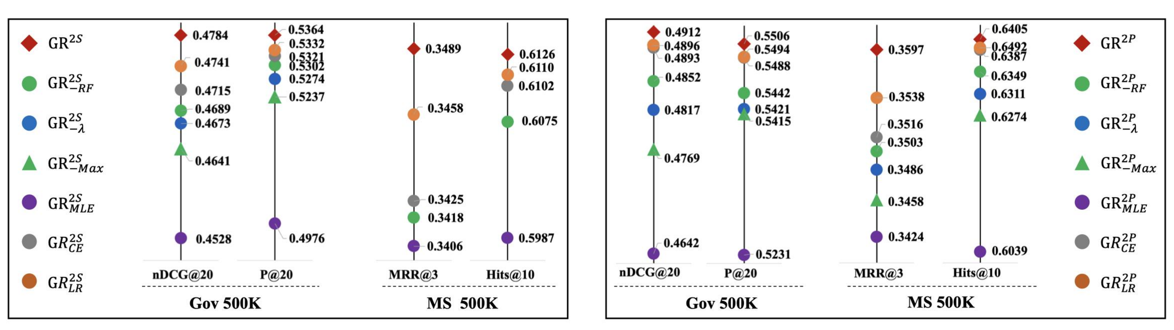


Figure 2: Ablation analysis. (Left) Supervised learning; (Right) Pre-training and fine-tuning.

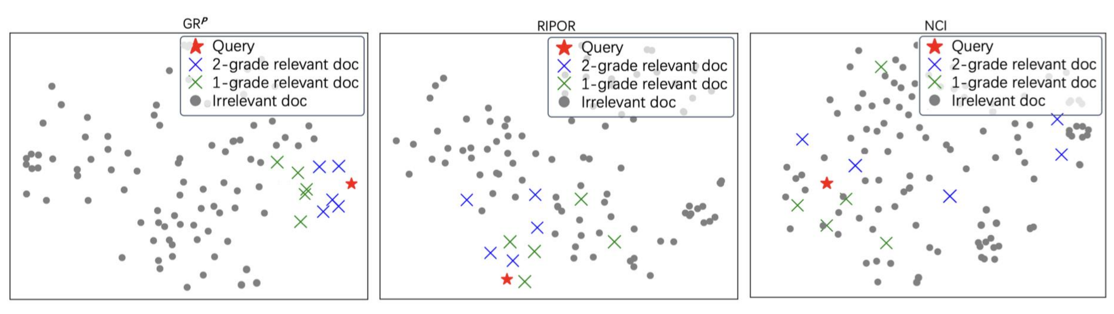


Figure 4: t-SNE plots of query and document representations for GR^{2P} (left), RIPOR (mid) and NCI (right).

Conclusion & Future work

- **Conclusion**:
  - We have proposed a MGCC loss for multi-graded GR that captures the relationships between multi-graded documents in a ranking, and a regularized fusion method to generate distinct and relevant docids. They work together to ensure more accurate GR retrieval. Empirical results on binary and multi-graded relevance datasets have demonstrated the effectiveness of the proposed method.
- **Future directions**:
  - We adopt hard weights for each relevance grade; what is the effect of a soft assignment setting in the MGCC loss?
  - The generated docids remain fixed after initialization; how to perform joint optimization of the docid generation and the retrieval task?

References

[1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. In International Conference on Learning Representations, 2022.

[2] Jeffrey S. Beis and David G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In Proceedings of IEEE computer society conference on computer vision and pattern recognition, pages 1000–1006, 1997.

[3] Michael Bendersky, Donald Metzler, and W Bruce Croft. Learning concept importance using a weighted dependence model. In Proceedings of the third ACM international conference on Web search and data mining, pages 31–40, 2010.

[4] Michael Bendersky, Donald Metzler, and W Bruce Croft. Parameterized concept weighting in verbose queries. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pages 605–614, 2011.

[5] Michael Bendersky, Donald Metzler, and W Bruce Croft. Effective query formulation with multiple information sources. In Proceedings of the fifth ACM international conference on Web search and data mining, pages 443–452, 2012.

[6] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9):509–517, 1975.

[7] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen-tau Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive search engines: Generating substrings as document identifiers. In Advances in Neural Information Processing Systems, 2022.

[8] Keping Bi, Qingyao Ai, and W Bruce Croft. Asking clarifying questions based on negative feedback in conversational search. In Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval, pages 157–166, 2021.

[9] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In Proceedings of the 22nd international conference on Machine learning, pages 89–96, 2005.

[10] Christopher JC Burges. From Ranknet to Lambdarank to Lambdamart: An overview. Learning, 11(23-581):81, 2010.

[11] Wei-Cheng Chang and Yu. Pre-training tasks for embedding-based large-scale retrieval. arXiv preprint arXiv:2002.03932, 2020.