A novel Relevance Score for Unsupervised Retrieval with Large Language Models

Andrea D'Angelo
andrea.dangelo@graduate.univaq.it
Università degli Studi dell'Aquila
L'Aquila, Italy

Giovanni Stilo
giovanni.stilo@univaq.it
Università degli Studi dell'Aquila
L'Aquila, Italy

Antinisca di Marco
antinisca.dimarco@univaq.it
Università degli Studi dell'Aquila
L'Aquila, Italy

ABSTRACT

Large Language Models (LLMs) have found widespread success in many Natural Language Processing (NLP) tasks. In particular, in unsupervised document retrieval and Retrieval Augmented Generators (RAGs), LLMs are typically employed by pooling their embeddings, resulting in a Relevance Score function defined as the dot product between the mean vectors of a query and a document. However, collapsing the term embeddings into a single sentence embedding may lead to a loss of valuable information, potentially reducing ranking effectiveness. This research proposes DbU-Cloud, a novel density-based method to address these challenges in unsupervised document ranking by eliminating pooling layers from the computation of the Relevance Score for each document, and instead considering a density-based metric derived from outlier detection.

CCS CONCEPTS

• **Information systems → Document representation**; **Retrieval models and ranking**.

KEYWORDS

Document Ranking, Information Retrieval, Large Language Models

ACM Reference Format:
Andrea D'Angelo, Giovanni Stilo, and Antinisca di Marco. 2018. A novel Relevance Score for Unsupervised Retrieval with Large Language Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large Language Models (LLMs), with OpenAI's GPT [8] and Google's BERT [3] being the most widely recognized, have revolutionized the field of Natural Language Processing (NLP)[12]. Their success has prompted their integration into nearly every downstream task, both supervised and unsupervised, encompassing fields like Information Retrieval (IR) [1] and Semantic Search [2]. Specifically, an emerging field of research is the one concerning Retrieval Augmented Generation (RAG), which involves integrating a generator model with an Document Ranking module. Employing LLMs for Document Ranking involves the pooling of term embeddings into a single sentence embedding, which can represent either a document or a query. With this approach, the scoring function is determined by measuring the similarity between the mean dense embeddings of the query and a document, commonly using metrics like cosine similarity or the inner dot product. However, collapsing the term embeddings by pooling them into a single sentence embedding might result in undesirable ranking results and degraded effectiveness.

To address these issues, we propose DbU-Cloud, a novel Unsupervised method for Document Ranking with LLMs. This method does not employ a pooling layer for the embeddings, opting instead to compute a density-based Query-Document Relevance Score on the respective sets of embeddings. DbU-Cloud rewards not only the proximity of the two sets but their relative density as well. Moreover, as an Unsupervised method, it does not require further fine-tuning of the employed LLMs.

2 RELATED WORK

The rapid development of Large Language Models caused a paradigm shift in Document Retrieval. Sentence-BERT[9] is the foundation of current state-of-the-art dense retrievers that improve upon BERT's ranking by adding a Pooling Layer (to obtain a sentence embedding from term embeddings) and then fine-tuning the LLM on the semantic similarity task.

Fine-tuned LLMs are able to obtain state-of-the-art results on specific downstream tasks with general-purpose datasets. Khattab and Zaharia [5], in their research work for ColBERT, share an idea that is close to our motivation: instead of collapsing the tokens' embeddings into one sentence tensor, we can obtain a more fine-grained (Query, document) relevance score by considering all embeddings individually. However, ColBERT needs to be re-trained, while DbU-Cloud is an unsupervised method of ranking.

There is increasing effort in recent research on enhancing Retrieval-Augmented Generation (RAG) frameworks for Large Language Models (LLMs) [4, 6]. These innovative models integrate a retrieval stage, which is critical for supplementing the model's response capabilities. While the majority of existing studies focus on optimizing post-retrieval or pre-retrieval techniques (such as Query Rewriting [7]), our approach distinguishes itself by refining both Query and Document representations in a model-agnostic way.

Regardless of the model's underlying architecture, the general wide-spread practice to score documents is to aggregate the dense embeddings of each sentence and query into a single vector and then compute their distance. There are multiple drawbacks to this practice. Tu et al. [11] showed that the representation quality of single fixed-width tensors decreases as text length increases. Lastly,

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

	ALL-MPNET				DistilRoBERTa				DPR			
	MaP	R@10	MRR	DCG@10	MaP	R@10	MRR	DCG@10	MaP	R@10	MRR	DCG@10
XoE	0.160	0.082	0.551	1.400	0.145	0.090	0.549	1.519	0.090	0.071	0.418	1.037
MoE	0.175	0.087	0.550	1.405	0.180	0.100	0.600	1.610	0.110	0.096	0.448	1.115
DbU-Cloud	**0.215**	**0.094**	**0.622**	**1.727**	**0.198**	**0.109**	**0.601**	**1.754**	**0.145**	**0.125**	**0.517**	**1.343**

Table 1: Analysis of the LLM's impact on the *MaP*, *R*@10, *MRR*, and *DCG*@10 of the LLM-based baselines (MoE, XoE, DbU-Cloud). The values are computed by the mean across all corpora.

the specific fine-tuning that LLMs undergo does not allow them to easily transfer knowledge to other contexts, as shown by Thakur et al. [10]. The paper shows that zero-shot domain adaptation for these networks is very low.

3 METHODOLOGY

In this section, we explain the planned Methodology and Experimental Settings. Figure 1 depicts the planned workflow for the experimental analysis.

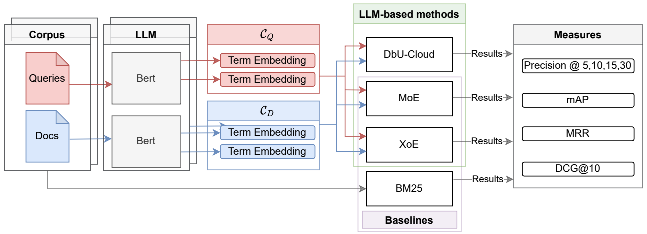

Figure 1: Experimental workflow planned for the experimental analysis. A pair of query/document is shown for clarity.

Query and documents are first transformed into sets of term embeddings (C_Q and C_D, respectively) via an LLM of choice (in Figure 1, BERT is used as an example). Then, we either apply the proposed method, or we apply mean pooling (MoE) or max pooling (XoE). The ranked lists produced by these methods are then evaluated with the most popular and widely recognized measures to evaluate their quality, such as Precision, mAP, MRR, and DCG.

3.1 DbU-Cloud

We define local density as in Equation 1:

$$\text{local-density}(\mathbf{e_i}, \mathbf{e_j}, C_D, k) = \min\left(\text{sim}(\mathbf{e_i}, \mathbf{e_j}), \text{akin}(\mathbf{e_j}, C_D, k)\right) \quad (1)$$

Where $akin(e_j, C_D, k)$ is the similarity between e_j and k-th most similar embedding in C_D. Moreover, we define the akin neighborhood $\mathcal{A}(\mathbf{e_i}, C_D, k)$ as the set of k embeddings closest to e_j:

$$\mathcal{A}(\mathbf{e_i}, C_D, k) = \{\mathbf{e_j} \in C_D \setminus \{\mathbf{e_i}\} \mid sim(\mathbf{e_i}, \mathbf{e_j}) \geq akin(\mathbf{e_i}, C_D, k)\} \quad (2)$$

Then, the DbU-Cloud scoring method that we use is shown in Equation 3:

$$DbU(C_Q, C_D, k) = \sum_{\mathbf{e_q} \in C_Q} \sum_{\mathbf{e_d} \in \mathcal{A}(\mathbf{e_q}, C_D, k)} \frac{\text{local-density}(\mathbf{e_q}, \mathbf{e_d}, C_D, k)}{|C_Q| \cdot |\mathcal{A}(\mathbf{e_q}, C_D, k)|} \quad (3)$$

DbU serves as a direct scoring function, meaning that its score can directly be used to compute the relevance score for a document D with respect to a query Q.

3.2 LLMs and Datasets

To evaluate the robustness of DbU-Cloud in different linguistic conditions, we purposefully chose corpora that differ in size, content, and language usage, varying from everyday language to specialized medical terminology. Namely, the chosen datasets were CISI, LISA[1], MS_MARCO[2] and NFCORPUS[3]. We also selected, from the literature, several LLMs with different underlying architectures and training methods, namely All-MPNET, DistilRoBERTa and DPR[4].

4 EARLY RESULTS

Table 1 shows early results for DbU-cloud, meaned across all corpora. The results of our analysis reveal that DbU-Cloud outperforms traditional pooling embeddings across all models evaluated and across all metrics. Specifically,the effectiveness of DbU-Cloud is particularly notable in improving the quality of the first retrieved results, as indicated by higher DCG scores, and also in enhancing the relevance of results deeper down the list, as shown by improved MaP scores. The effectiveness of DbU-Cloud suggests that its approach to handling embeddings is more adept at capturing and utilizing relevant information for retrieval tasks. These findings underscore the robustness and efficiency of DbU-Cloud, making it a preferable choice for improving the accuracy and reliability of Document retrieval systems.

5 FUTURE WORK

Future work for DbU-Cloud includes extending the analysis to other baselines (specifically the non-LLM based baseline of BM25), other models, and other corpora. We also aim to formalize the mathematical framework in which DbU-Cloud operates to strengthen its theoretical foundations. Lastly, we would like to explore some empirical examples on specific queries to understand its strengths and pitfalls.

ACKNOWLEDGMENTS

All the numerical simulations have been realized on the Linux HPC cluster Caliban of the High-Performance Computing Laboratory of the Department of Information Engineering, Computer Science and Mathematics (DISIM) at the University of L'Aquila.

[1] CISI and LISA are available at https://ir.dcs.gla.ac.uk/resources/test_collections/
[2] MS_MARCO is available at https://microsoft.github.io/msmarco/
[3] The NFCORPUS is available at https://www.cl.uni-heidelberg.de/statnlpgroup/nfcorpus/
[4] All the models we employed are available from the hugging face repository.

REFERENCES

[1] Giambattista Amati. 2009. *Information Retrieval.* Springer US, Boston, MA, 1519–1523. https://doi.org/10.1007/978-0-387-39940-9_915

[2] Philippe Cudre-Mauroux. 2018. *Semantic Search.* Springer International Publishing, Cham, 1–6. https://doi.org/10.1007/978-3-319-63962-8_231-1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[4] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL]

[5] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[7] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query Rewriting for Retrieval-Augmented Large Language Models. arXiv:2305.14283 [cs.CL]

[8] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training.

[9] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[10] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. arXiv:2104.08663 [cs.IR]

[11] Zhengkai Tu, Wei Yang, Zihang Fu, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2020. Approximate Nearest Neighbor Search and Lightweight Dense Vector Reranking in Multi-Stage Retrieval Architectures. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (Virtual Event, Norway) *(ICTIR '20)*. Association for Computing Machinery, New York, NY, USA, 97–100. https://doi.org/10.1145/3409256.3409818

[12] Karin Verspoor and Kevin Bretonnel Cohen. 2013. *Natural Language Processing.* Springer New York, New York, NY, 1495–1498. https://doi.org/10.1007/978-1-4419-9863-7_158

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009