Incorporating Cognitive Complexity of Text in Dense Retrieval for Personalized Search

Effrosyni Sokli[1,*], Alessandro Raganato[1] and Gabriella Pasi[1]

[1]Università degli Studi di Milano-Bicocca, Dipartimento di Informatica, Sistemistica e Comunicazione DISCo, Viale Sarca 336, Milano, Italy

Abstract

Personalized search aims to tailor the retrieval results to the particular interests of individual users. The user's cognitive level and level of expertise in specific domains are some of the key factors to be considered during the development of personalized models. This work proposes the integration of a Mixture-of-Experts architecture with Dense Retrieval Models, leveraging the cognitive complexity levels defined in Bloom's Taxonomy for personalized Information Retrieval based on the user's level of expertise in educational search.

Keywords

Dense Retrieval, Mixture-of-Experts, Cognitive Complexity

1. Introduction

Personalization in Information Retrieval (IR) aims to tailor the search outcome based on the user's (or group of users') preferences and their online behavior [1, 2, 3]. In this paper, we address personalized IR from the aspect of the user's cognitive complexity level on a given field, as defined in Bloom's Taxonomy [4]. This newly considered factor can be beneficial for specific personalization tasks, such as in an educational search scenario, where the search results can be personalized based on information about the user's level of expertise on a given topic. Since neural models were widely adopted in IR, the development and application of Dense Retrieval Models (DRMs) [5, 6], which leverage deep learning to learn semantic relationships between queries and documents (semantic embeddings) [7], significantly increased. DRMs achieve high performance in domain-specific search scenarios [8], where IR systems are designed for specialized fields or subjects [9], and given their great adaptability to various tasks [7], there is also the potential for their application in personalized IR [10].

This work proposes an approach that integrates a Mixture-of-Experts (MoE) architecture with DRMs to leverage and incorporate the cognitive complexity levels defined in Bloom's Taxonomy [11], aiming to enhance the generated embedding representations of text and provide personalized search outcomes based on the user's level of expertise in educational search.

2. Background and Related Work

Bloom's Taxonomy. For the proposed work, we adopt the Cognitive Complexity of text as defined in Bloom's Taxonomy [11]. Bloom's Taxonomy, introduced in 1956 and later revised (Revised Bloom's Taxonomy - RBT) in the early 2000s [4], aims to facilitate educators to optimize the conveyance of knowledge by classifying learning objectives into six different levels of increasing cognitive complexity, i.e., Remember, Understand, Apply, Analyze, Evaluate, Create. This classification of text into different levels of cognitive complexity intends to implicitly depict the user's level of expertise on the given topic. Figure 1 illustrates some examples of sentences classified into the RBT categories with incrementing cognitive complexity. While the readability of the text is not considered for this classification, RBT

IIR2024: 14th Italian Information Retrieval Workshop, 5th and 6th September 2024, Udine, Italy
*Corresponding author.
effrosyni.sokli@unimib.it (E. Sokli); alessandro.raganato@unimib.it (A. Raganato); gabriella.pasi@unimib.it (G. Pasi)
0009-0003-5388-2385 (E. Sokli); 0000-0002-7018-7515 (A. Raganato); 0000-0002-6080-8170 (G. Pasi)

Learning Objective Examples	Labels
Recognise the key role that human factors play in the leadership and development of a highly functional perioperative team.	Remember
Describe the general characteristics of the modern X-ray system used in clinical practice, including scientific principles, and production of the digital image.	Understand
Apply research skills to operate effectively as a member of a research project team.	Apply
Identify an issue of relevance to the practice of perioperative medicine capable of further investigation and research within the context of a capstone project.	Analyze
Ability to articulate critical interpretations of dramatic texts and processes in systematic written argument.	Evaluate
A capacity to design, manage, and carry out a research project.	Create
Analyze and apply contemporary management theory and research to current organizational issues.	Apply & Analysis
Assess and synthesise diverse information about up-to-date information and knowledge management systems market and how to use implementation strategies to maximise their strengths and minimise their weaknesses.	Evaluate & Create

Figure 1: Examples of sentences classified into the six classes of the Revised Bloom's Taxonomy (RBT).

focuses on how knowledgeable the user is on the given topic. RBT consists of two dimensions - Knowledge and Cognitive, with the latter being widely used in the literature for various tasks of Text Classification in levels of cognitive complexity [12, 13, 14].

Mixture-of-Experts. MoE [15] is a Machine Learning (ML) technique used in various contexts within the ML community [16], and also in the field of NLP [17, 18, 19]. MoE allows the division of complex problems into sub-domains or sub-tasks, each managed by a uniquely trained expert, increasing the underlying model's adaptivity. This technique has been effectively applied to enhance Transformer-based dense models across various NLP tasks by incorporating linguistic knowledge, such as sentence structural complexity [20, 21]. In this work, we describe our plan to inject linguistic knowledge about the cognitive complexity of text by integrating a MoE architecture in a DRM.

3. Proposed Methodology

Motivations. Bloom's Taxonomy is designed to help educators create meaningful learning objectives for their courses by classifying them into gradually increasing levels of cognitive complexity. Therefore, an appropriate application for our proposed model is in an Educational Search Engine scenario for personalized retrieval based on the user's educational level on a given course or topic. Furthermore, the cognitive complexity of a text as defined in Bloom's Taxonomy can be also associated with the user's expertise on the related given topic. Hence, by taking into account user data on tasks such as domain-specific retrieval, we consider the employment of the proposed model for personalization. To give an example, in an academic environment where two distinct students are known to have or have not attended a certain course (e.g., physics), the search outcome for the same (physics-related) query should differ based on their course attendance data. More specifically, the model should retrieve documents with higher cognitive complexity for the student who has attended a given topic considering that they would be able to comprehend more advanced related content.

The Proposed Model. The method of adapter tuning [22, 23] is employed for the implementation of the MoE architecture. The proposed model builds upon a traditional bi-encoder DRM architecture, enhancing it with an MoE module that targets both query and document representations. A bi-encoder architecture [5, 24] allows for independent encoding of documents and queries, which enhances scalability and enables the computation of relevance scores through the similarity function of the query and document representations. A Transformer-based network is commonly used to embed both queries and documents. While it is possible to employ separate encoders [25], using a single encoder for both queries and documents has been shown to improve robustness, without significantly affecting performances [5, 26].

Each expert is trained in a different cognitive complexity class (i.e., Remember, Understand, Apply, Analyze, Evaluate, and Create as defined in RBT) leading to six experts in total for the proposed model. The MoE component receives the input embedding from the tokenizer and applies a series of transformations through specialized experts (Figure 2). The output results in six modified embedding representations, enriched with expert representations. The representations are then weighted by the gating function and aggregated by the pooling module to form the final embedding representation, which is eventually used for similarity assessment.

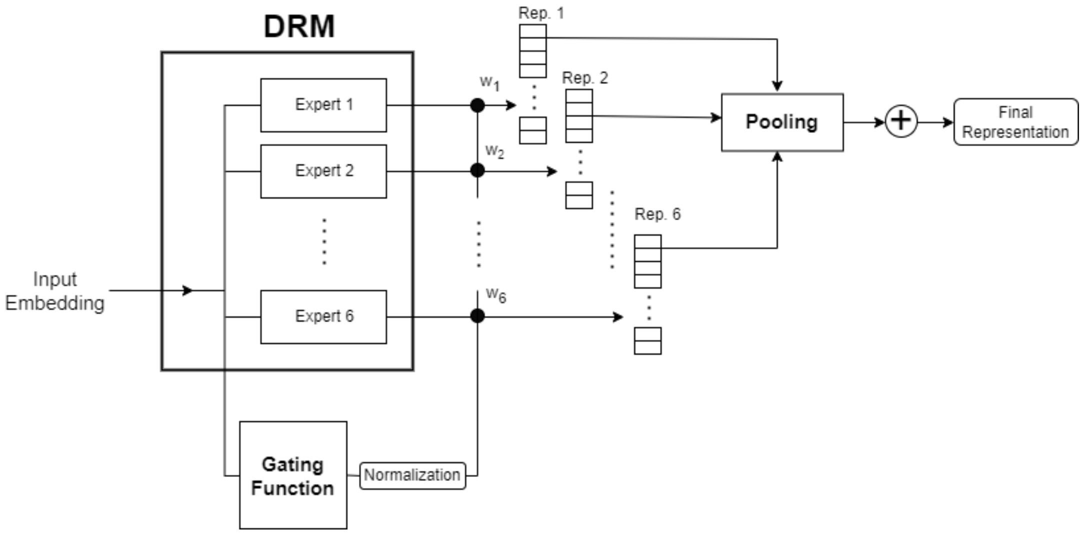

Figure 2: The overall architecture of the proposed model.

The gating function computes the likelihood for the input text to belong to any of the six predefined RBT classes. This mechanism relies on a six-label classifier and does not assume the mutual exclusivity of labels. Instead, it allows an input to belong to multiple RBT classes. Finally, the pooling module merges the enhanced representations computed by the experts on the basis of the RBT class likelihood, estimated by the gating function in the form of a normalized vector of six weights. Merging is achieved by simply weighting and summing up the outputs of the experts, as shown in Figure 2.

4. Planned Experiments & Evaluation Strategy

In order to gain meaningful insights concerning the model's performance in various tasks and domains, we consider the following two standard IR benchmarks.

Zero-shot Evaluation. As an initial step, we consider the evaluation of our model in a zero-shot retrieval scenario. Specifically, we plan on evaluating the proposed approach on the BEIR benchmark collection (BEnchmarking IR [27]), a heterogeneous benchmark for zero-shot evaluation of IR models, allowing us to assess our approach without prior user and task-specific training. With this setting, we aim to evaluate the adaptability of the proposed model across various tasks and scenarios and showcase the proposed approach's effects on the robustness and retrieval performance of the underlying dense retriever. At this evaluation phase, the model will be assessed as a first-stage dense retriever and will be compared with the best-performing dense retrievers on the BEIR benchmark (e.g., Contriever [28]) on metrics such as nDGC@10 and Recall@100 [29].

Multi-Domain Benchmark for Personalized Search Evaluation. Moreover, we consider evaluating the proposed approach on the task of personalized domain-specific academic search, by exploiting the evaluation benchmark proposed by Bassani et al. [30]. The authors provided a multi-domain evaluation benchmark for Personalized Academic Search with more than 18 million documents and 1.9 million queries across four domains (i.e., computer science, physics, political science, and psychology), along with baseline performance including pre-trained neural models, especially for the evaluation of

Personalized Re-Ranking models. This setting enables the evaluation of the model's ability to enhance the underlying DRM's embedding representations with expert insights, thereby personalizing search outcomes during the re-ranking stage according to the user's expertise. At this stage, the proposed model will be evaluated as a Personalized Re-Ranking model, with retrieval performance assessed based on user data reflecting their expertise on a given topic.

Limitations. Besides the aforementioned evaluation approaches and personalization strategies in IR being under study for a long time, there is still a lack of high-quality large-scale benchmark datasets for conducting offline comparative evaluations, especially in an educational search scenario, which constitutes one of the main limitations that have to be addressed in the future.

5. Conclusions & Future Work

This work proposes an approach that incorporates information about the cognitive complexity of text in DRMs using an MoE architecture, to generate enhanced embedding representations for personalized search based on the user's level of expertise on a given topic in educational search. Our model leverages Bloom's Taxonomy, a method developed to classify text into different levels of increasing cognitive complexity, to inject linguistic knowledge into DRMs and improve retrieval performance. Initially, we plan the evaluation of our approach in a zero-shot retrieval scenario, without any prior user or task-specific training, to conduct an overall assessment of the model's retrieval performance and capabilities. Consequently, we intend extending the application and evaluation of our approach to the task of personalized domain-specific academic search by leveraging available benchmarks in the literature. In this scenario, we intend assessing the model's capabilities in personalized search based on the user's level of expertise on a given topic. Although at this stage the model would personalize the search results based solely on the user's level of expertise, we consider expanding the personalization model by taking into account an enriched user profile with additional information about the user's preferences and prior searches as future work. Ultimately, the proposed approach considers a novel dimension for personalized IR models and aims to create space for further research in the field of educational search by taking into account the user's cognitive level and level of expertise in a certain field.

Acknowledgments

This work has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101073307.

References

[1] A. Borisov, I. Markov, M. de Rijke, P. Serdyukov, A context-aware time model for web search, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 205–214. URL: https://doi-org.unimib.idm.oclc.org/10.1145/2911451.2911504. doi:10.1145/2911451.2911504.

[2] M. Braga, A. Raganato, G. Pasi, Personalization in bert with adapter modules and topic modelling, in: Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023). Pisa, Italy, 2023, pp. 24–29.

[3] S. Calegari, G. Pasi, Personal ontologies: Generation of user profiles based on the yago ontology, Information Processing & Management 49 (2013) 640–658. URL: https://www.sciencedirect.com/science/article/pii/S0306457312001070. doi:https://doi.org/10.1016/j.ipm.2012.07.010, personalization and Recommendation in Information Access.

[4] D. R. Krathwohl, A revision of bloom's taxonomy: An overview, Theory into practice 41 (2002) 212–218.

[5] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, Transactions on Machine Learning Research (2022). URL: https://openreview.net/forum?id=jKN1pXi7b0.
[6] Y. Yu, C. Xiong, S. Sun, C. Zhang, A. Overwijk, COCO-DR: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1462–1479. URL: https://aclanthology.org/2022.emnlp-main.95. doi:10.18653/v1/2022.emnlp-main.95.
[7] B. Mitra, N. Craswell, An introduction to neural information retrieval, Foundations and Trends® in Information Retrieval 13 (2018) 1–126. URL: http://dx.doi.org/10.1561/1500000061. doi:10.1561/1500000061.
[8] R. Pradeep, Y. Li, Y. Wang, J. Lin, Neural query synthesis and domain-specific ranking templates for multi-stage clinical trial matching, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 2325–2330. URL: https://doi-org.unimib.idm.oclc.org/10.1145/3477495.3531853. doi:10.1145/3477495.3531853.
[9] M. Lupu, M. Salampasis, A. Hanbury, Domain specific search, Professional search in the modern world: Cost action IC1002 on multilingual and multifaceted interactive information access (2014) 96–117.
[10] H. Zeng, S. Kallumadi, Z. Alibadi, R. Nogueira, H. Zamani, A personalized dense retrieval framework for unified information access, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 121–130. URL: https://doi-org.unimib.idm.oclc.org/10.1145/3539618.3591626. doi:10.1145/3539618.3591626.
[11] B. S. Bloom, Taxonomy of Educational Objectives: Handbook II, David McKay, 1956.
[12] Y. Li, M. Rakovic, B. Poh, D. Gasevic, G. Chen, Automatic classification of learning objectives based on bloom's taxonomy, in: A. Mitrovic, N. Bosch (Eds.), Proceedings of the 15th International Conference on Educational Data Mining, International Educational Data Mining Society, 2022, pp. 530–537. URL: https://educationaldatamining.org/edm2022/,https://educationaldatamining.org/edm2022/proceedings/. doi:10.5281/zenodo.6853191, educational Data Mining 2022, EDM 2022 ; Conference date: 24-07-2022 Through 27-07-2022.
[13] M. Mohammed, N. Omar, Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec, PLOS ONE 15 (2020) 1–21. URL: https://doi.org/10.1371/journal.pone.0230442. doi:10.1371/journal.pone.0230442.
[14] A. A. Yahya, A. Osman, Automatic classification of questions into bloom's cognitive levels using support vector machines (2011).
[15] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive Mixtures of Local Experts, Neural Computation 3 (1991) 79–87. URL: https://doi.org/10.1162/neco.1991.3.1.79. doi:10.1162/neco.1991.3.1.79.
[16] S. Masoudnia, R. Ebrahimpour, Mixture of experts: a literature survey, Artificial Intelligence Review 42 (2014) 275–293.
[17] B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, W. Fedus, St-moe: Designing stable and transferable sparse expert models, arXiv preprint arXiv:2202.08906 (2022).
[18] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. P. Bosma, Z. Zhou, T. Wang, E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. Le, Y. Wu, Z. Chen, C. Cui, GLaM: Efficient scaling of language models with mixture-of-experts, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 5547–5569. URL: https://proceedings.mlr.press/v162/du22c.html.
[19] W. Fedus, B. Zoph, N. Shazeer, Switch transformers: Scaling to trillion parameter models with

simple and efficient sparsity, Journal of Machine Learning Research 23 (2022) 1–39.

[20] A. Miaschi, F. Dell'Orletta, G. Venturi, Linguistic knowledge can enhance encoder-decoder models (if you let it), 2024. arXiv:2402.17608.

[21] W. Wang, B. Bi, M. Yan, C. Wu, J. Xia, Z. Bao, L. Peng, L. Si, Structbert: Incorporating language structures into pre-training for deep language understanding, in: International Conference on Learning Representations, 2019.

[22] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 2790–2799. URL: https://proceedings.mlr.press/v97/houlsby19a.html.

[23] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, I. Gurevych, AdapterFusion: Non-destructive task composition for transfer learning, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 487–503. URL: https://aclanthology.org/2021.eacl-main.39. doi:10.18653/v1/2021.eacl-main.39.

[24] P. Kasela, G. Pasi, R. Perego, N. Tonellotto, Desire-me: Domain-enhanced supervised information retrieval using mixture-of-experts, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 111–125.

[25] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550. doi:10.18653/v1/2020.emnlp-main.550.

[26] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), 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, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[27] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021. URL: https://openreview.net/forum?id=wCu6T5xFjeJ.

[28] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, Transactions on Machine Learning Research (2024).

[29] E. Kamalloo, N. Thakur, C. Lassance, X. Ma, J.-H. Yang, J. Lin, Resources for brewing beir: Reproducible reference models and statistical analyses, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1431–1440. URL: https://doi-org.unimib.idm.oclc.org/10.1145/3626772.3657862. doi:10.1145/3626772.3657862.

[30] E. Bassani, P. Kasela, A. Raganato, G. Pasi, A multi-domain benchmark for personalized search evaluation, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3822–3827. URL: https://doi-org.unimib.idm.oclc.org/10.1145/3511808.3557536. doi:10.1145/3511808.3557536.