Axiomatic Guidance for Efficient and Controlled Neural Search

Andrew Parry
a.parry.1@research.gla.ac.uk
University of Glasgow
Glasgow, UK

ABSTRACT

Pre-trained language models based on the transformer architecture [13], provide solutions to general ad-hoc search tasks–ranging from news search to question-answering–vastly outperforming statistical approaches in terms of both precision and recall [9, 16]. These models operate over "semantics", removing the need for bespoke features based on proprietary data (e.g., interaction logs). In doing so, this paradigm may lead to further adoption of the idealised "end-to-end" retrieval system as an elegant and powerful search solution. However, outside of sanitised benchmarks, these models present exploitable and untrustworthy biases [8, 10] relinquishing any control over inference due to their black-box nature.

Definition 1 (Bias). *Biases are factors in neural estimation of relevance which were unintended by system design*

Such biases threaten the viability of neural models in production. Without greater control over model output, stakeholders could raise concerns hindering the adoption of effective and efficient search. Today, feature-based search systems are still performant relative to state-of-the-art neural search and can adapt to a changing corpus and the needs of system stakeholders. As agency over information access is further reduced via emerging paradigms such as Retrieval-Augmented-Generation [4, 5], we must retain control over the output of a search system. We posit that by allowing external features to influence the semantic interactions within neural search at inference time, as illustrated in Figure 1, we can not only allow control over system output but reduce the need to model corpus-specific priors, which can instead be modelled by external features allowing for greater generalisation and training efficiency gains.

We consider that bias in neural search systems is an artefact of the training and underlying mechanisms of current pre-trained models but is not present in statistical models. Features such as statistical models are principled [1, 2] and arbitrarily controllable; these features can adapt to a corpus and meet the demands of a given search task. Conversely, the output of a current neural system can only be changed by post hoc constraints [3] or by re-training the underlying model. Additionally, training models that can outperform classical approaches out-of-domain frequently requires multi-model negative mining [14], multi-stage distillation [16], and increasingly large models [11, 15]. This process, now cited as common practice [7, 12], is expensive and removes any ability to attribute performance to a particular part of training or inference components. Resolving the dichotomy between the term mismatch of the explainable and controlled statistical model and the flexibility of the biased neural model is a challenging problem. Nevertheless, we propose that by taking principles from axiomatic approaches, we can rectify biases in neural search whilst improving efficiency. As presented in Figure 1, we aim to reduce the complexity of neural ranker training and inference, applying classical IR principles during training as a generalisable process as opposed to the ad-hoc constraint of prior work [3, 6]. Axiomatic signals can guide and control neural ranking models to reduce spurious factors in semantic relevance estimation by compensating for the frozen priors of neural systems whilst still operating over flexible latent space. Given the biases observed in current systems, this may satiate the concerns of multiple stakeholders, leading to broader adoption of the paradigm.

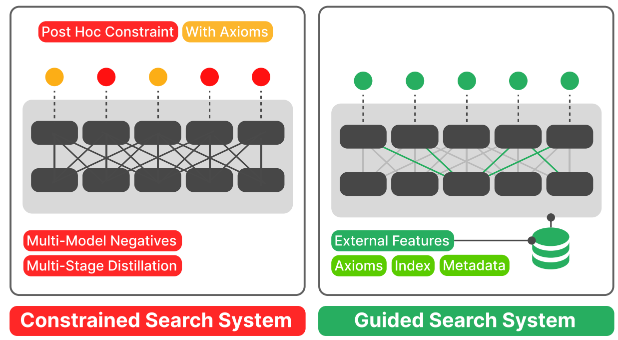

Figure 1: Proposed benefits of Axiomatic Guidance in Search, primarily we look to move away from a data-driven approach with post-hoc constraint to enable efficient deployment of neural systems with greater control over the output of a semantic model.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Axiomatic Retrieval, Neural ranking, Interpretability, Efficiency

ACM Reference Format:
Andrew Parry. 2024. Axiomatic Guidance for Efficient and Controlled Neural Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3626772.3657651

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657651

REFERENCES

[1] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://api.semanticscholar.org/CorpusID:15200693

[2] Hui Fang, Tao Tao, and Chengxiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29, 2, Article 7 (apr 2011), 42 pages. https://doi.org/10.1145/1961209.1961210

[3] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Association for Computational Linguistics, 3030–3042. https://doi.org/10.18653/V1/2021.NAACL-MAIN.241

[4] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics, 6769–6781. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.550

[5] Patrick S. H. 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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[6] Minghan Li, Sheng-Chieh Lin, Barlas Oguz, Asish Ghoshal, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. CITADEL: Conditional Token Interaction via Dynamic Lexical Routing for Efficient and Effective Multi-Vector Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 11891–11907. https://doi.org/10.18653/V1/2023.ACL-LONG.663

[7] Guangyuan Ma, Xing Wu, Zijia Lin, and Songlin Hu. 2024. Drop your Decoder: Pre-training with Bag-of-Word Prediction for Dense Passage Retrieval. arXiv:2401.11248 [cs.IR]

[8] Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the Behavior of Neural IR Models. *Transactions of the Association for Computational Linguistics* 10 (2022), 224–239. https://doi.org/10.1162/tacl_a_00457

[9] Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *CoRR* abs/1901.04085 (2019). arXiv:1901.04085 http://arxiv.org/abs/1901.04085

[10] Andrew Parry, Maik Fröbe, Sean MacAvaney, Martin Potthast, and Matthias Hagen. 2024. Analyzing Adversarial Attacks on Sequence-to-Sequence Relevance Models. In *Advances in Information Retrieval*. Springer Nature Switzerland, Cham, 286–302.

[11] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! *CoRR* abs/2312.02724 (2023). https://doi.org/10.48550/ARXIV.2312.02724 arXiv:2312.02724

[12] Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Xiaolong Huang, Binxing Jiao, Linjun Yang, and Daxin Jiang. 2023. LexMAE: Lexicon-Bottlenecked Pretraining for Large-Scale Retrieval. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=PfpEtB3-csK

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[14] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2023. SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 2244–2258. https://doi.org/10.18653/v1/2023.acl-long.125

[15] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving Text Embeddings with Large Language Models. arXiv:2401.00368 [cs.CL]

[16] Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics, 538–548. https://doi.org/10.18653/V1/2022.EMNLP-MAIN.35

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