EqualizeIR: Mitigating Linguistic Biases in Retrieval Models

Jiali Cheng **Hadi Amiri**
University of Massachusetts Lowell
{jiali_cheng, hadi_amiri}@uml.edu

Abstract

This study finds that existing information retrieval (IR) models show significant biases based on the linguistic complexity of input queries, performing well on linguistically simpler (or more complex) queries while underperforming on linguistically more complex (or simpler) queries. To address this issue, we propose EqualizeIR, a framework to mitigate linguistic biases in IR models. EqualizeIR uses a *linguistically biased* weak learner to capture linguistic biases in IR datasets and then trains a robust model by regularizing and refining its predictions using the biased weak learner. This approach effectively prevents the robust model from overfitting to specific linguistic patterns in data. We propose four approaches for developing linguistically-biased models. Extensive experiments on several datasets show that our method reduces performance disparities across linguistically simple and complex queries, while improving overall retrieval performance.

1 Introduction

Neural ranking models have been extensively used in information retrieval and question answering tasks (Dai and Callan, 2020; Zhao et al., 2021; Khattab and Zaharia, 2020; Karpukhin et al., 2020; Xiong et al., 2021; Hofstätter et al., 2021). We demonstrate that these models can show strong linguistic biases, where the retrieval performance is biased with respect to the “linguistic complexity” of queries, quantified by the variability and sophistication in productive vocabulary and grammatical structures in queries using existing tools (Lu, 2010, 2012; Lee et al., 2021; Lee and Lee, 2023).[1]

Figure 1 shows that the average linguistic complexity of the test queries in the NFCorpus (Boteva et al., 2016) and FIQA (Maia et al., 2018) datasets varies significantly, where the NDCG@10 performance of the BM25 model significantly decreases on NFCorpus and improves on FIQA as the linguistic complexity of queries increase. This performance disparity across queries of different linguistic complexity leads to the focus of this paper and the following research question: *can we debias IR models to achieve equitable performance across queries of varying linguistic complexity?*

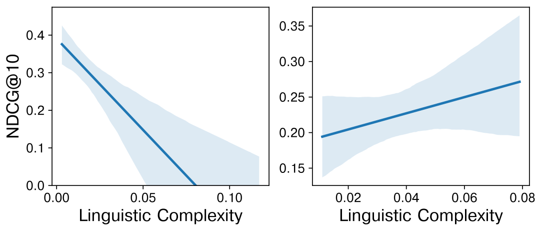

Figure 1: NDCG@10 of BM25 on the test set of NFCorpus (Boteva et al., 2016) (left) decreases and on the test set of FIQA (Maia et al., 2018) (right) increases as the average linguistic complexity (Lu, 2010, 2012) of queries increase. Specifically, we observe a significant drop in NDCG@10, from 0.4 to 0, and a significant increase in NDCG@10, from 0.2 to 0.3. The result shows that BM25 is significantly biased toward linguistically easy and hard examples on different datasets.

Inspired by previous debiasing works in natural language processing (Utama et al., 2020; Ghaddar et al., 2021; Sanh et al., 2021; Meissner et al., 2022), we introduce a new approach, named EqualizeIR, to mitigate linguistic biases in IR models. EqualizeIR is a *weak learner* framework; it first trains a linguistically-biased weak learner to explicitly capture linguistic biases in a dataset. This linguistically-biased weak learner is then used as a reference to inform and regularize the training of a desired (robust) IR model. It encourages the IR model to focus less on biased patterns and more on the underlying relevance signals. This is achieved by using the biased weak learner’s predictions as indicators of bias intensity in inputs, and adjusting the IR model’s predictions accordingly.

[1] We consider lexical and syntactic linguistic complexity indicators in this study. Details of these indicators are provided in Appendix B, Table 4.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 889–898
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

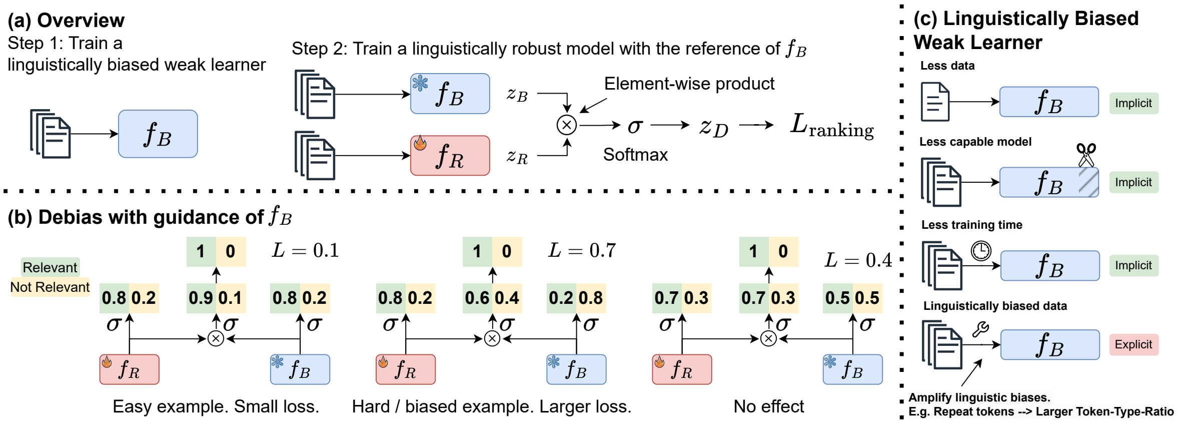
(a) Overview
Step 1: Train a linguistically biased weak learner
f_B
Step 2: Train a linguistically robust model with the reference of f_B
f_B
f_R
z_B
z_R
Element-wise product
$\sigma \longrightarrow z_D \longrightarrow L_{\text{ranking}}$
Softmax
(c) Linguistically Biased Weak Learner
Less data
f_B
Implicit
Less capable model
f_B
Implicit
Less training time
f_B
Implicit
Linguistically biased data
f_B
Explicit
Amplify linguistic biases.
E.g. Repeat tokens --> Larger Token-Type-Ratio
(b) Debias with guidance of f_B
Relevant
Not Relevant
1 0
$L = 0.1$
0.8 0.2
0.9 0.1
0.8 0.2
σ
f_R
f_B
Easy example. Small loss.
1 0
$L = 0.7$
0.8 0.2
0.6 0.4
0.2 0.8
σ
f_R
f_B
Hard / biased example. Larger loss.
1 0
$L = 0.4$
0.7 0.3
0.7 0.3
0.5 0.5
σ
f_R
f_B
No effect

Figure 2: Architecture of EqualizeIR for mitigating linguistic biases in IR models. (a) Training process: first, a linguistically biased IR model f_B is trained. Then, we freeze the parameters of f_B to train a target, linguistically robust IR model f_R by taking the product of logits of f_B and f_R. The biased weak learner regularizes the ranking loss of f_R using its learned linguistic biases. (b): Examples showing that the ensemble approach effectively moderates prediction probabilities to avoid learning biases associated with high confidence or moving too heavily toward the biased weak learner. (c): Strategies for developing linguistically biased weak learners.

EqualizeIR does not require linguistic biases to be explicitly described for the model, and reduces the risk of overfitting to specific types of biases. Specifically, we investigate several strategies to develop a linguistically-biased weak learner: training the model using **linguistically biased data** to directly introduce and reinforce specific linguistic patterns, using a **weaker model** with fewer parameters or a simpler architecture to reduce models ability to generalize across inputs with various linguistic complexity, **shortening the training time** to prevent the model from capturing the diversity and depth of linguistic features in the data, and training on a **limited data** to emphasize the linguistic features present in a specific subset of data. Through these strategies, we aim to develop a model that effectively captures linguistic biases for developing linguistically robust IR models.

Our contribution are (a): illustrating that the performance of current IR models vary based on the linguistic complexity of input queries, (b): a novel approach that trains a linguistically robust IR model with the help of a linguistically biased IR model to mitigate such biases, and (c): four approaches to obtain linguistically biased weak learners, all effective in mitigating biases in IR models.

2 EqualizeIR

Linguistic Complexity: measures sophistication in productive vocabulary and grammatical structures in textual content, spanning lexical, syntactic, and discourse dimensions. In this work, we adopt existing linguistic complexity measurements (lexical complexity (Lu, 2012) and syntactic complexity (Lu, 2010)) to measure the linguistic complexity of queries in IR datasets implemented by existing tools (Lu, 2010, 2012; Lee et al., 2021; Lee and Lee, 2023). Specifically, given a query q, a linguistic complexity score is computed by averaging scores of various linguistic complexity metrics, which includes measures such as verb sophistication and the number of T-units. The detailed list of linguistic complexity is shown in Appendix B Table 4. We column normalize linguistic complexity scores before computing average linguistic complexity for each query.

Overview: EqualizeIR mitigates linguistic biases in an IR model using a linguistically-biased weak learner, f_B. The process begins with training f_B to learn linguistic biases present in a dataset. Then, a linguistically robust model, f_R, is trained based on the confidence of f_B (which approximates the intensity of linguistic biases in input) and the prediction accuracy of f_R. This approach has two purposes: firstly, f_B guides f_R to improve its robustness by learning from the identified biases of f_B. Secondly, f_B can adjust the weights of training examples by prioritizing those that f_R fails to predict, which effectively refines the training focus of f_R toward more challenging examples.

Bi-Encoder Architecture: We consider a standard bi-encoder architecture with a query encoder f_q and a document encoder f_d (Khattab and Zaharia, 2020; Karpukhin et al., 2020; Xiong et al., 2021; Hofstätter et al., 2021). Given the i-th batch $\mathcal{B}_i = \{q_i, d_i^+, d_{i,1}^-, \dots, d_{i,n}^-\}$, where q_i denotes the query, d_i^+ denotes a relevant document, and $d_{i,j}^-, \forall j$ denote irrelevant documents, we encode them into embeddings $h_{q_i}, h_{d_i^+}, h_{d_i^-}$, and optimize the standard contrastive loss:

$$L = -\log \frac{e^{\mathrm{sim}(h_q, h_{d^+})}}{e^{\mathrm{sim}(h_q, h_{d^+})} + \sum_{j=1}^{n} e^{\mathrm{sim}(h_q, h_{d_j^-})}} \quad (1)$$

2.1 Debiasing with Biased Weak Learner

We first train a linguistically biased weak learner f_B using the bi-encoder architecture to model dataset biases. After training, we freeze f_B's parameters and use it to train f_R. Given an input example $x_i = (q_i, d_i)$, we first obtain the logits from the linguistically-biased weak learner f_B and the target linguistically robust model f_R:

$$z_B = f_B(x_i), \quad z_R = f_R(x_i). \quad (2)$$

As Figure 2(a) shows, to integrate the knowledge from the linguistically biased weak learner into the training of the target IR model f_R, we compute the element-wise product of the two probabilities and normalize it with a softmax function, or more conveniently element-wise addition in log spcae:

$$\log(z_D) = \sigma\big(\alpha \log(z_B) + \log(z_R)\big), \quad (3)$$

where $\alpha \in [0, 1]$ is a scaling factor that controls the strength of the effect of the biases detected by f_B on the final output of f_R. This adjusted probability z_D is the debiased probability (see the rationale below), which is then used to compute a standard ranking loss, where f_B remains frozen and only the parameters of f_R are updated. This approach encourages f_R to adopt a less linguistically biased stance under the guidance of f_B.

We note that the effect of element-wise product can be interpreted from two perspectives: (a): dynamic curriculum: here the importance of training samples within a batch are adaptively re-weighted based on the confidence of f_B's prediction; and (b): regularization function: here f_B act as regularizer by constraining f_R to avoid excessive confidence in its predictions, particularly for easy samples that it already predicts correctly. Consequently, f_R does not overfit to specific biased patterns within the dataset. Therefore f_B acts as both a guide and guard to make f_R a more robust model against linguistic bias.

This approach effectively refines the training of f_R using the weak learner f_B. Figure 2(b) provides several examples of the functionality of f_B. In case (1), when f_B confidently makes a correct prediction, f_R is adjusted to increase its confident in the correct label, as the input is likely an easy example. This lowers the loss (compared to f_R's actual loss), reduces the weight of the example in training of f_R, and effectively minimizes the risk of learning biases from the example by f_R. In case (2), when f_B confidently makes a wrong prediction, it indicates that the input sample likely contains biases that mislead f_B. Here, f_R's confidence is adjusted to learn from the example by generating a larger than original loss, which encourages the model to adapt to these hard samples.

2.2 Strategies for Developing Biased Learners

Previous findings show that a "weak" model learns and relies on superficial patterns for making predictions (Utama et al., 2020; Ghaddar et al., 2021; Sanh et al., 2021; Meissner et al., 2022). We introduce four approaches to obtain a linguistically-biased weak learner (f_B) from both model and data perspectives.

- First, we obtain a biased weak learner by **repeating linguistic constructs**, such as noun phrases, in queries. This approach makes the model more sensitive to complex linguistic structures by amplifying them in queries without changing the semantics.
- Second, we train a **weaker model** with limited capacity to learn complex patterns, making it weaker in terms of predictive power but useful for exposing biases. This weaker model can be either a completely separate model (e.g. TinyBERT (Turc et al., 2019)) or a subset of f_R (Cheng and Amiri, 2024).
- Third, we use the same architecture as the target IR model, but train it with significantly **fewer iterations**, which results in an "undercooked" version that is weaker.
- Finally, we train the model on **less data**, which reduces its ability to generalize and learn deeper patterns.

Each of these weak learners reveal different linguistic biases in data, and provide insights into the biases that f_R needs to overcome. Appendix 4, Figure 5 shows that the above approaches indeed result in linguistically biased f_Bs.

3 Experiments

Datasets We use the *test* sets of four IR datasets form BEIR benchmark (Thakur et al., 2021):

- **MS MARCO** (Nguyen et al., 2016), a passage retrieval dataset with 532k training samples and 43 test queries;
- **NFCorpus** (Boteva et al., 2016), a biomedical IR dataset with 110k training samples and 323 test queries,
- **FIQA-2018** (Maia et al., 2018), a question answering dataset with 14k training samples and 648 test queries, and
- **SciFact** (Wadden et al., 2020), a scientific fact checking dataset with 920 training samples and 300 test queries.

IR Models We compare our approach to the following baselines:

- **BM25** (Robertson et al., 2009; Manning, 2009), which retrieves documents based on lexical similarity;
- **DPR** (Karpukhin et al., 2020), a dense retrieval model that compute similarity in embedding space;
- **ColBERT** (Khattab and Zaharia, 2020), which adopts a delayed and deep interaction of token embeddings of query and document;
- **Multiview** (Amiri et al., 2021), a multiview IR approach with data fusion and attention strategies;
- **RankT5** (Zhuang et al., 2023), the Seq2Seq model (Raffel et al., 2023);
- **KernelWhitening** (Gao et al., 2022), which learns sentence embeddings that disentangles causal and spurious features; and
- **LC as Rev Weight**, which uses linguistic complexity to reversely weight the probability.

Evaluation Following previous works (Thakur et al., 2021; Zhuang et al., 2023), we use NDCG@10 as the evaluation metric. We report average ($\mu, \uparrow$), standard deviation ($\sigma, \downarrow$), and coefficient of variation ($c_v = \frac{\sigma}{\mu}, \downarrow$) of NDCG@10 across all test queries. In addition, we examine models' performance in terms of the linguistic complexity of test examples. A robust model should have high overall performance and low performance variation across the spectrum of linguistic complexity (e.g. easy, medium, hard). Due to the limited space, we only implement EqualizeIR to DPR.

Method	$\mu(\uparrow)$	$\sigma(\downarrow)$	$c_v(\downarrow)$
BM25	0.44	0.32	0.82
ColBERT	0.29	0.43	1.71
DPR	0.29	0.32	1.23
RankT5	0.42	0.25	0.64
Multiview	0.42	0.26	0.66
KernelWhitening	0.44	0.25	0.57
LC as Rev Weight	0.27	0.21	0.78
EqualizeIR	**0.47**	**0.22**	**0.52**

Table 1: Main results. μ, σ, and c_v denote average performance, standard deviation, and coefficient of variation across test queries. Best performance is in **bold** and second best is underlined. The significance test is shown in Table 3.

4 Main Results

Existing IR models are linguistically biased Figure 3 and Table 1 show that existing IR models are linguistically biased with significant performance fluctuations as the linguistic complexity of query increases, resulting in a disparate performance across different levels of linguistic complexity. On average, BM25, DPR, ColBERT, RanKT5, and Multiview have varied performance across queries, with high standard deviation of 0.32, 0.32, 0.43, 0.25 and 0.26. These results highlight the need to mitigate linguistic biases in these models.

EqualizeIR increases average performance and reduces linguistic bias EqualizeIR outperforms BM25, DPR, ColBERT, RankT5, and Multiview by 0.03, 0.15, 0.15, 0.05, and 0.05 absolute points in average NDCG@10 respectively, while also showing smaller standard deviation in NDCG@10 across all test queries. EqualizeIR outperforms baselines in terms of c_v (NDCG@10) by large margins of 0.30, 0.71, 1.19, 0.08, and 0.14 compared to BM25, DPR, ColBERT, RankT5, Multiview respectively.

Different IR models show different linguistic biases On NFCorpus, BM25 achieves 0.40 NDCG@10 on linguistically easy examples, while close to zero NDCG@10 on hard examples. Conversely, DPR perform poorly on linguistically easy examples and better on linguistically hard examples. This contrasting results can be attributed to the underlying architectures of the IR models, such as the text encoders and if late interaction is used, and the intrinsic characteristics of the datasets.

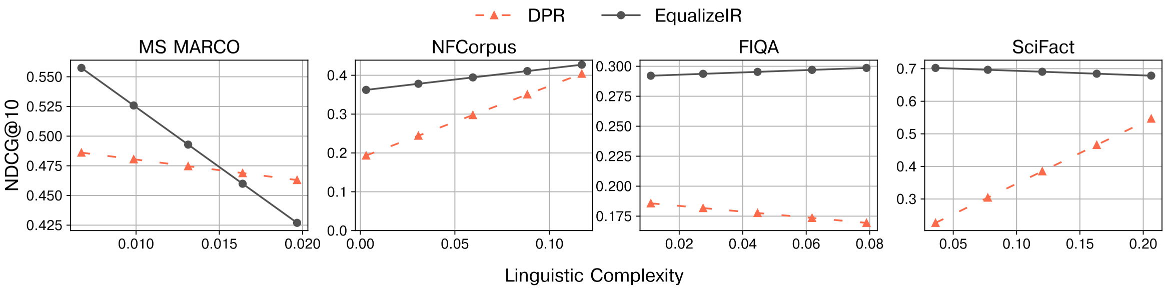

Figure 3: NDCG@10 of EqualizeIR and DPR (Karpukhin et al., 2020) as linguistic complexity of queries increase. Detailed performance of all baselines is shown in Figure 4 in Appendix A.

Comparison Between Different Biased Models Figure 5 shows that, as we hypothesized, all four types of weak learners encode substantial linguistic biases. Results in Appendix A Table 5-8 show the comparison between different methods to obtain f_B. Overall, different f_B training methods have similar overall performance and performance variation in terms of NDCG@10. We notice that that the "weaker model" and "less data" approaches consistently yield higher NDCG@10 performance, which may indicate that they better capture linguistic biases for f_R to avoid. In contrast, the "repeating linguistic constructs" and "fewer iterations" strategies do not produce a good biased learner. This result could be attributed to the models potential overemphasis on specific linguistic features or lack of learning discriminative patterns from data, while overshadowing other aspects that may contribute to bias and resulting in a less effective bias detection. In addition, the "weaker model" and "less data" approaches may capture a broader type of biases, including implicit ones, which makes them more flexible and practical. Using a less capable model as f_B leads to the highest overall performance, smallest performance deviation and variation. Using less data has a slightly lower overall performance and higher performance deviation. This comparison highlights that different f_Bs exhibit different linguistic biases and result in varying performances of f_R.

5 Related Work

Information Retrieval DPR (Karpukhin et al., 2020) and ColBERT (Khattab and Zaharia, 2020) are earlier works of dense retrieval, where similarity is computed in high-dimensional embedding space. Although effective, Faggioli et al. (2024) prove that operating in query-specific subspaces can improve the performance and efficiency of dense retrieval models. Recently, more attention has been paid to adapting Large Language Models (LLMs) to information retrieval (Guo et al., 2024; Xu et al., 2024; Borges et al., 2024).

Bias Mitigation Li et al. (2022) design an in-batch regularization technique to mitigate the biased performance across different subgroups. Kim et al. (2024) propose to identify semantically relevant query-document pairs to explain why documents are retrieved, and discover that existing IR models show biased performances across different brand name. Ziems et al. (2024) discover that IR models suffer from indexical bias, i.e. the bias resulted by the order of documents, and propose a new metric DUO to evaluate the amount of indexical bias an IR model has. Query performance prediction (QPP) (Arabzadeh et al., 2024) studies whether we can predict the IR quality by only looking at the query itself without additional information. On other tasks, prior works have discussed how biased models or weak learners can be applied to debiasing in vision (Cadene et al., 2019), natural language understanding (Sanh et al., 2021; Ghaddar et al., 2021; Cheng and Amiri, 2024), and speech classification tasks (Cheng et al., 2024).

6 Conclusion

We report that IR models are biased toward linguistic complexity of queries and introduce EqualizeIR, a framework that trains a robust IR model by regularizing it with four types of linguistically-biased weak learners (by amplifying linguistic constructs in queries, using a weaker model with limited capacity, training with fewer iterations to create an underdeveloped model, and training on less data to restrict generalization), to achieve equitable performance across queries of varying linguistic complexity.

Limitations

Existing definitions of linguistic complexity often have a narrow focus on specific linguistic features, which can result in challenges in comprehensive quantification of linguistic biases. For example, we did not consider linguistic biases related to discourse, pragmatics, morphology and semantics. In addition, our debiasing approach slightly increases complexity of training by requiring a trained biased model. Similar to other debiasing approaches, there's a risk of model overfitting to particular biases the model is trained to address, which may limit its adaptability to generalize to new or unseen biases. Finally, although our approach can be applied to any supervised IR model, we only applied it dense retrieval models, and its performance on other IR models remained underexplored.

Broader Impact Statement

We present an important issue in existing IR models: they show disparate and biased performance across queries with different levels of linguistic complexity–quantified by lexical and syntactic complexity. This can disproportionately disadvantage queries from users with specific writing style that result in particular types of linguistic complexity. It is important that future research and evaluation protocols in IR accounts for these biases and mitigate them.

References

Hadi Amiri, Mitra Mohtarami, and Isaac Kohane. 2021. Attentive multiview text representation for differential diagnosis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1012–1019, Online. Association for Computational Linguistics.

Negar Arabzadeh, Chuan Meng, Mohammad Aliannejadi, and Ebrahim Bagheri. 2024. Query performance prediction: From fundamentals to advanced techniques. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V*, page 381–388, Berlin, Heidelberg. Springer-Verlag.

Luís Borges, Rohan Jha, Jamie Callan, and Bruno Martins. 2024. Generalizable tip-of-the-tongue retrieval with llm re-ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2437–2441, New York, NY, USA. Association for Computing Machinery.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval*, pages 716–722, Cham. Springer International Publishing.

Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jiali Cheng and Hadi Amiri. 2024. FairFlow: Mitigating dataset biases through undecided learning for natural language understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21960–21975, Miami, Florida, USA. Association for Computational Linguistics.

Jiali Cheng, Mohamed Elgaar, Nidhi Vakil, and Hadi Amiri. 2024. Cognivoice: Multimodal and multilingual fusion networks for mild cognitive impairment assessment from spontaneous speech. In *Interspeech 2024*, pages 4308–4312.

Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1533–1536, 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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Guglielmo Faggioli, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. 2024. Dimension importance estimation for dense information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1318–1328, New York, NY, USA. Association for Computing Machinery.

SongYang Gao, Shihan Dou, Qi Zhang, and Xuanjing Huang. 2022. Kernel-whitening: Overcome dataset bias with isotropic sentence embedding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4112–4122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abbas Ghaddar, Phillippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. 2021. End-to-end self-debiasing framework for robust NLU training.

In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1923–1929, Online. Association for Computational Linguistics.

Ping Guo, Yubing Ren, Yue Hu, Yanan Cao, Yunpeng Li, and Heyan Huang. 2024. Steering large language models for cross-lingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 585–596, New York, NY, USA. Association for Computing Machinery.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

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. *arXiv preprint arXiv:2004.04906*.

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*, pages 39–48.

Youngwoo Kim, Razieh Rahimi, and James Allan. 2024. Discovering biases in information retrieval models using relevance thesaurus as global explanation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19530–19547, Miami, Florida, USA. Association for Computational Linguistics.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Yuantong Li, Xiaokai Wei, Zijian Wang, Shen Wang, Parminder Bhatia, Xiaofei Ma, and Andrew Arnold. 2022. Debiasing neural retrieval via in-batch balancing regularization. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 58–66, Seattle, Washington. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Christopher D Manning. 2009. *An introduction to information retrieval*. Cambridge university press.

Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2022. Debiasing masks: A new framework for shortcut mitigation in NLU. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7607–7613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

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

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. 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)*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Shuyuan Xu, Wenyue Hua, and Yongfeng Zhang. 2024. Openp5: An open-source platform for developing, training, and evaluating llm-based recommender systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 386–394, New York, NY, USA. Association for Computing Machinery.

Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575, Online. Association for Computational Linguistics.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2308–2313, New York, NY, USA. Association for Computing Machinery.

Caleb Ziems, William Held, Jane Dwivedi-Yu, and Diyi Yang. 2024. Measuring and addressing indexical bias in information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12860–12877, Bangkok, Thailand. Association for Computational Linguistics.

A Addition Results

We present the performance with respect to linguistic complexity in Figure 4 and the performance on each dataset in Table 2. Overall, the results show that existing IR models are linguistically biased, showing significant performance fluctuations as the linguistic complexity of query changes. Table 5-8 compares the performances between different methods to obtain f_B.

Data	Method	$\mu(\uparrow)$	$\sigma(\downarrow)$	$c_v\,(\downarrow)$
FIQA	BM25	0.25	0.32	1.26
	ColBERT	0.23	0.22	0.96
	DPR	0.22	**0.18**	0.82
	RankT5	0.26	0.21	0.81
	Multiview	0.27	0.23	0.85
	EqualizeIR	**0.29**	0.21	**0.72**
MS MARCO	BM25	**0.48**	0.25	0.53
	ColBERT	0.44	0.38	0.88
	DPR	0.47	0.29	0.61
	RankT5	0.43	0.33	0.77
	Multiview	0.42	0.35	0.83
	EqualizeIR	**0.48**	**0.20**	**0.42**
NFCorpus	BM25	0.34	0.32	0.92
	ColBERT	0.28	0.25	0.89
	DPR	0.31	0.27	0.87
	RankT5	0.33	0.29	0.88
	Multiview	0.32	0.28	0.88
	EqualizeIR	**0.37**	**0.23**	**0.62**
SciFact	BM25	0.69	0.39	0.56
	ColBERT	0.50	0.34	0.68
	DPR	0.40	0.32	0.80
	RankT5	0.68	0.33	0.49
	Multiview	0.64	0.36	0.56
	EqualizeIR	**0.70**	**0.25**	**0.36**

Table 2: Main results. μ, σ, and c_v denote average performance, standard deviation, and coefficient of variation across all queries in each test set. Best performance is in **bold** and second best is underlined.

Data	MS MARCO	NFCorpus	FIQA	SciFact
BM25	2.8e-3	9.0e-4	2.2e-3	2.8e-2
ColBERT	1.5e-3	2.0e-9	2.9e-12	1.1e-36
DPR	2.9e-3	1.4e-13	3.3e-13	9.3e-38
RankT5	1.1e-3	1.7e-4	9.1e-5	1.1e-2
Multiview	1.4e-5	6.0e-14	8.1e-14	1.4e-8

Table 3: Significance test between EqualizeIR and baselines adjusted with bonferroni correction. Results show that EqualizeIR performs significantly better than baselines.

B Linguistic Complexity

Table 4 presents the 45 linguistic complexity measurements in our study. For the full description of these metrics, see (Lu, 2010, 2012; Lee and Lee, 2023). We provide a brief description of a few indices as an example: **Type–Token Ratio, TTR** is the ratio of unique words in the text. **D-measure** is a modification to TTR that accounts for text length. * **Variation** indicates variations in lexical words such as nouns, verbs, adjectives, and adverbs. The **Mean Length of T-Units** is the average length of T-units in text. A T-unit is defined as a minimal terminable unit, essentially an independent clause and all its subordinate clauses. It provides insight into the syntactic complexity by measuring how elaborate the clauses are on average.

Type	Index Name	Notation
Syntactic	Mean length of clause	MLC
	Mean length of sentence	MLS
	Mean length of T-Unit	MLT
	Sentence complexity ratio	C/S
	T-unit complexity ratio	C/T
	Complex T-unit proportion	CT/T
	Dependent Clause proportion	DC/C
	Dependent Clause to T-Unit ratio	DC/T
	Sentence coordination ratio	T/S
	Coordinate phrases to clause ratio	CP/C
	Coordinate phrases to T-Unit ratio	CP/T
	Complex nominals to clause ratio	CN/C
	Complex nominals to T-unit ratio	CN/T
	Verb phrases to T-unit ratio	VP/T
Lexical	Type–Token Ratio TTR	T/N
	Mean TTR of all 50-word segments	MSTTR–50
	Corrected TTR CTTR	T/$\sqrt{2N}$
	Root TTR RTTR	T/$\sqrt{N}$
	Bilogarithmic TTR	$\log(TTR)$ $\log(T)$ / $\log(N)$
	Uber Index Uber	$\log(2N)$ / $\log(N/T)$
	D Measure	D
	Lexical Word Variation	LV Tlex/Nlex
	Verb Variation-I	VV1 T_{Verb} / N_{Verb}
	Squared VV1	SVV1 Tv2
	Verb	N_{Verb}
	Corrected VV1	CVV1 T_{Verb} /$\sqrt{2Nverb}$
	Verb Variation-II	T_{Verb} /Nlex
	Noun Variation	T_{Noun} / Nlex
	Adjective Variation	AdjV T_{Adj} /Nlex
	Adverb Variation	AdvV T_{Adv} /Nlex
	Modifier Variation	ModV (T_{Adj} + T_{Adv})/ Nlex

Table 4: Linguistic indices used in the study

Dataset	$\mu(\uparrow)$	$\sigma(\downarrow)$	$c_v(\downarrow)$
Less data	0.27	0.23	0.85
Less capable model	**0.29**	**0.21**	**0.72**
Less trained	0.27	0.24	0.89
Linguistically biased data	0.26	0.26	1.01

Table 5: Comparison of different strategies for developing linguistically biased models in terms of NDCG@10 on FIQA. Best performance is in **bold** and second best is underlined.

C Implementation Details

We use PyTorch (Paszke et al., 2019) and BEIR (Thakur et al., 2021) to implement our approach. For DPR and ColBERT, we use BERT-base (Devlin et al., 2019) as the encoders. For f_B trained with less data, we randomly take 20% of the original training data to train f_B. For f_B trained with less capable model, we use BERT-Tiny (Turc et al., 2019) as the encoder. For f_B

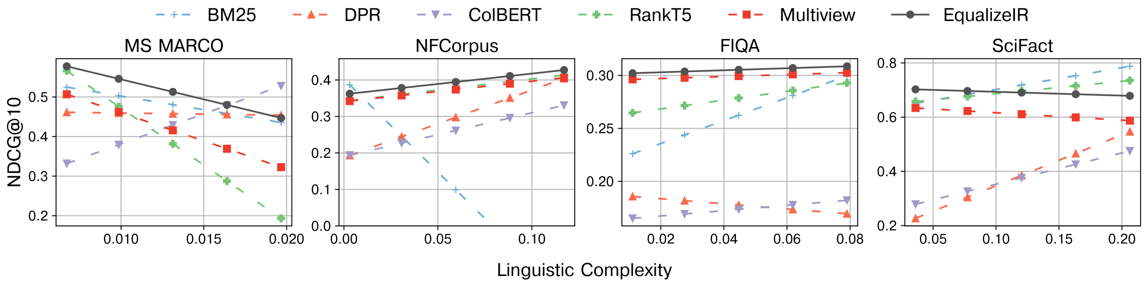

Figure 4: Performance in NDCG@10 as linguistic complexity of queries increase.

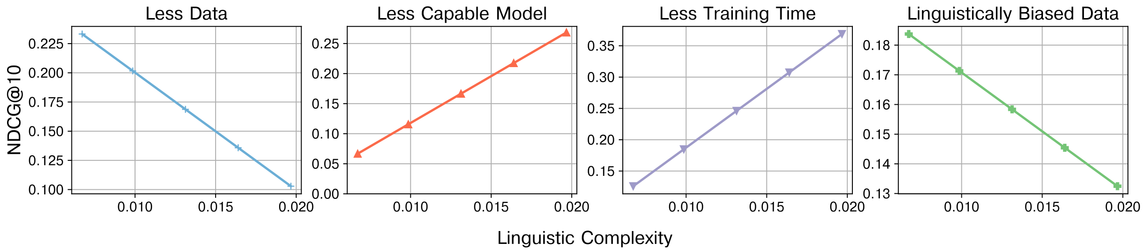

Figure 5: Performance of f_B obtained by four different strategies, which are highly linguistically biased.

Dataset	$\mu(\uparrow)$	$\sigma(\downarrow)$	$c_v(\downarrow)$
Less data	0.44	0.23	0.52
Less capable model	**0.48**	**0.20**	**0.42**
Less trained	0.42	0.26	0.62
Linguistically biased data	0.42	0.25	0.60

Table 6: Comparison of different strategies for developing linguistically biased models in terms of NDCG@10 on MS MARCO. Best performance is in **bold** and second best is underlined.

Dataset	$\mu(\uparrow)$	$\sigma(\downarrow)$	$c_v(\downarrow)$
Less data	0.33	0.27	0.81
Less capable model	**0.37**	**0.23**	**0.62**
Less trained	0.35	0.25	0.71
Linguistically biased data	0.32	0.26	0.81

Table 7: Comparison of different strategies for developing linguistically biased models in terms of NDCG@10 on NFCorpus. Best performance is in **bold** and second best is underlined.

trained with less time, we train it for 20% of the original training time. All methods are trained with AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of $1e-5$. We tune α on validation sets and find choosing $\alpha = 0.1$ yields best performance consisitently across datasets.

Dataset	$\mu(\uparrow)$	$\sigma(\downarrow)$	$c_v(\downarrow)$
Less data	0.68	0.33	0.49
Less capable model	**0.70**	**0.25**	**0.36**
Less trained	0.67	0.35	0.52
Linguistically biased data	0.61	0.30	0.49

Table 8: Comparison of different strategies for developing linguistically biased models in terms of NDCG@10 on SciFact. Best performance is in **bold** and second best is underlined.