Causalign: Using Intervention-based Regularization to Improve Out-of-Distribution Performance of Long-Form Text Matching Systems

Daniel Frees [*1,2] **Martin Pollack** [*1]

1. Topic

Recommendation systems are critical to web browsing and information retrieval, powering systems ranging from Netflix TV show recommendations to modern LLM-based retrieval augmented generation (RAG). Much of the data enabling these systems takes the form of unstructured text, thus motivating the problem of *text-based matching*. Text-based matching is fundamentally a challenge in determining matching signal between sparsely related queries and items. This sparsity, combined with the fact that spurious correlations due to noise are context-dependent (e.g. a brand such as 'Google' might be semantically relevant for matching in one scenario, and spurious in another) makes text matching very difficult. Recent research has begun to investigate whether causal methodologies such as intervention-based regularization might be employed to better separate semantically relevant text from spurious correlations, thereby improving the results of text-based recommendation systems as data distributions change. This increased robustness is especially necessary for systems such as web searches where both query (search) and item (webpage) distributions are frequently changing.

2. Literature Review

Each of the following papers seeks to improve recommendation system generalization performance by reducing the effects of spurious correlations.

First, (Bansal et al., 2023) investigate a causal method for regularizing fine-tuned encoder models to restrain word-level causal effects to be similar to the base encoder model. Bansal et al. find that fine-tuned recommendation encoders perform worse than base models on out-of-distribution (OOD) data such as new item categories/ queries, and introduce an intervention-based regularizer to restrain the causal effect of individual words in a fine-tuned model to be similar to the `BERT` (or similar) base encoder. Their results show that causal based regularization can reduce the deleterious effects of fine-tuning and are recommendable for mild to moderate distribution shifts in the deployment of a recommendation system based on short text matching.

(Pang et al., 2021) approach the more challenging problem of long-form text matching, but similarly aim to minimize spurious noise. In contrast to (Bansal et al., 2023), their approach is algorithmic, using PageRank to filter noisy sentences and words in the attention blocks of a Transformer-based encoder model. Pang et al.'s results demonstrate that isolating semantically relevant inputs can improve not only OOD performance, but overall text-matching performance. These results suggest that causal intervention-based regularization methods are likely to be effective in the long-form text matching domain as well.

(Zhang et al., 2021) investigate the separation of causal and spurious components in encoder systems similarly to (Bansal et al., 2023), but for the task of sequential user interactions prediction. Zhang et al. learn user encoder f_θ and item encoder g_θ to predict a user u's interaction $y_{u,t}$ at time t based on their historical interaction sequence $x_{u,t} = y_{u,[:t]}$. Zhang et al. identify 'indispensable' item-level concepts by scoring the semantic item-encoding (g_θ) similarity between items throughout a user's interaction history, filter out the bottom half of 'dispensable' items (heuristically identified as noise), and then use contrastive learning with synthetic counterfactuals to ensure that users are pushed further apart by differences in 'indispensable' item history as compared with differences in 'dispensable' item history. Their results outperform state-of-the-art (SOTA) sequential recommenders, demonstrating that methods for isolation of causal components to denoise encodings can provide benefits for other kinds of recommender systems, not just text-matching systems. Further, their work suggests future directions for text-matching; it may be interesting to investigate whether similar contrastive learning on 'indispensible' text components can regularize text query embeddings.

3. Proposal

In this work, we seek to extend (Bansal et al., 2023)'s experiments to the problem of long-form text matching. As described in the literature review, (Bansal et al., 2023) showed how various "causal regularizers" or "intervention-based regularizers" for large language models (LLMs) can improve the out-of-distribution generalization of fine-tuned text-matching systems. Bansal et al. noted that fine-tuning massively increases performance on within-distribution data (queries and items from the same distribution the model was fine-tuned on), but actually decreases out-of-distribution performance compared to base pre-trained encodings. By employing regularization (`ITVReg`) to constrain changes in the causal effect of individual words, Bansal et al. were able to achieve superior F1 scores on out-of-distribution product recommendations data (e.g. new product categories). We aim to identify whether the deleterious effects of fine-tuning extend to the more complex problem space of long-form text matching, and to replicate Bansal's `ITVReg` method for long text matching. We will replicate Bansal et al.'s experiments comparing standard contrastive loss fine-tuning against causal-regularized fine-tuning for long-form text matching encoders, specifically focusing on the task of matching Association for Computational Linguistics (ACL) papers to their cited papers using paper abstracts (Radev et al., 2022).

[*]Equal contribution [1]Department of Statistics, Stanford University [2]Google, San Francisco, CA. Correspondence to: Daniel Frees <dfrees@stanford.edu>, Martin Pollack <pollackm@stanford.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

References

Bansal, P., Prabhu, Y., Kiciman, E., and Sharma, A. Using interventions to improve out-of-distribution generalization of text-matching recommendation systems, 2023. URL https://arxiv.org/abs/2210.10636.

Pang, L., Lan, Y., and Cheng, X. Match-ignition: Plugging pagerank into transformer for long-form text matching. *arXiv preprint arXiv:2101.06423*, 2021. URL https://arxiv.org/abs/2101.06423. Last revised 17 Aug 2021 (v2).

Radev, D. R., Muthukrishnan, P., Qazvinian, V., and Abu-Jbara, A. The acl anthology network corpus - huggingface curation. *Language Resources and Evaluation*, 47(4):919–944, 2022. doi: 10.1007/s10579-012-9211-2. URL https://huggingface.co/datasets/WINGNUS/ACL-OCL. Updated to September 2022.

Zhang, S., Yao, D., Zhao, Z., Chua, T.-s., and Wu, F. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, pp. 11, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8037-9. doi: 10.1145/3404835.3462908. URL https://doi.org/10.1145/3404835.3462908.