A Boring-yet-effective Approach for the Product Ranking Task of the Amazon KDD Cup 2022

Vitor Jeronymo
NeuralMind, Brazil
University of Campinas, Brazil

Guilherme Rosa
NeuralMind, Brazil
University of Campinas, Brazil

Surya Kallumadi
Lowes, USA

Roberto Lotufo
NeuralMind, Brazil
University of Campinas, Brazil

Rodrigo Nogueira
NeuralMind, Brazil
University of Campinas, Brazil

ABSTRACT

In this work we describe our submission to the product ranking task of the Amazon KDD Cup 2022. We rely on a receipt that showed to be effective in previous competitions: we focus our efforts towards efficiently training and deploying large language models, such as mT5, while reducing to a minimum the number of task-specific adaptations. Despite the simplicity of our approach, our best model was less than 0.004 nDCG@20 below the top submission. As the top 20 teams achieved an nDCG@20 close to 0.90, we argue that we need more difficult e-Commerce evaluation datasets to discriminate retrieval methods.

CCS CONCEPTS

• **Information systems** → **Online shopping**; **Specialized information retrieval**;

KEYWORDS

eCommerce information retrieval, product search, large language models, mT5

ACM Reference Format:
Vitor Jeronymo, Guilherme Rosa, Surya Kallumadi, Roberto Lotufo, and Rodrigo Nogueira. 2022. A Boring-yet-effective Approach for the Product Ranking Task of the Amazon KDD Cup 2022. In *KDD Cup 2022 Workshop: ESCI Challenge for Improving Product Search, August 17, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/

1 INTRODUCTION

Recent improvements in information retrieval, mainly due to pre-trained transformer models, opened up the possibility of improving search in various domains [2, 4, 5, 7–10, 12]. Among such domains, e-commerce search receives special attention by the industry as improvements in search quality often lead to increases in revenue.

In this work, we detail our submission to the Amazon KDD Cup 2022, whose goal is to evaluate ranking methods that can be used to improve the customer experience when searching for products.

KDDCup '22, August 17, 2022, Washington, DC, USA

2 RELATED WORK

Our solution is based on the monoT5 model, that demonstrated strong effectiveness in various passage ranking tasks in different domains. We qualify our method as "boring", since it is well known in the recent IR literature that models with more parameters can outperform smaller ones with task-specific adaptations. For example, Nogueira et al. [11] used the model to achieve state-of-the-art results on TREC 2004 Robust Track [21] while Pradeep et al. [13] used the same model, finetuned only on MS MARCO, to achieve the best or second best performance on medical domain ranking datasets, such as Precision Medicine [16] and TREC-COVID [23]. In addition, Rosa et al. [18, 19] used large versions of monoT5 to reach the state of the art in a legal domain entailment task in the COLIEE competition [6, 14]. Furthermore, Rosa et al. [17] showed that the 3 billion-parameter variant of monoT5 achieves the state of the art in 12 out of 18 datasets of the Benchmark-IR (BEIR) [20], which consists of datasets from different domains such as web, biomedical, scientific, financial and news.

3 METHODOLOGY

In this section, we describe mMonoT5, a multilingual variant of monoT5 [11], which is an adaptation of the T5 model [15] for the passage ranking task. We first finetune a multilingual T5 model [22] on the mMARCO dataset [3], which is the translated version of MS MARCO [1] in 9 languages. The model is trained to generate a "yes" or "no" token depending on the relevance of a document to a query.

mMonoT5 uses the following input template:

$$\text{Query: } q \quad \text{Document: } d \quad \text{Relevant:} \tag{1}$$

where q represents a query and d represents a document that may or may not be relevant to the given query.

During inference, the model receives the same input prompt and estimates a score s that quantifies the relevance of a document d to a query q by applying a softmax function to the logits of the "yes" and "no" tokens, and then taking the probability of the "yes" token as the final score. That is,

$$s = P(\text{Relevant} = 1 | d, q). \tag{2}$$

After computing all scores for a given query, we rank then with respect to their scores.

After finetuning on mMARCO, we further finetuned the model on the training data of tasks 1 and 2 of the competition. We use the

Model	nDCG@20 Public	Private
monoT5-3B (dataset translated to En)	0.8750	-
mMonoT5-580M (mMARCO only)	0.8640	-
mMonoT5-580M	0.8900	-
mMonoT5-3.7B (our best submission)	0.9012	0.9007
First place (team www)	0.9057	0.9043
20th place (team we666)	0.8933	0.8929

Table 1: Main results of the competition.

Beautiful Soup library to clean any remaining HTML tags that may appear in the product. Products are presented to the model as the concatenation of the fields product_title, product_description, product_bullet_point, product_brand and product_color_name, joined by whitespaces.

During the competition we observed that using task 2 training data improved the model substantially. Hence, we used task 1 and 2 training data by transforming the labeled data classes to "true" if 'exact' and all other classes as "false". We use these tokens instead of "yes" and "no", used by the original mMonoT5. We trained the model for 5 epochs, which takes about 72 hours in a TPU v3, using batches of 128 and maximum sequence length of 512 tokens.

4 RESULTS

We show our results in Table 1. Our best model achieved an nDCG@20 of 0.9012 and 0.9007 on the public and private test sets, respectively, placing us in the ninth place on the leaderboard and only 0.0036 behind the first position.

Initially, we used the mMonoT5 base, with 580M parameters, finetuned on mMarco data to test the model's zero-shot capability. This model achieves an nDCG@20 of 0.864. Then we further finetuned it on the training data of the competition, which results in a nDCG@20 of 0.89, which later, the 3.7B parameter version surpassed by 0.0112 points. We also tried translating the corpus and queries into English and using the monoT5-3B (English-only) finetuned on the competition data, but it could not out-do its multilingual counterpart.

5 CONCLUSION

We described a boring but effective approach based on the multilingual variation of monoT5 that achieved competitive results in the product ranking task of the Amazon KDD Cup 2022.

REFERENCES

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenber, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading Comprehension Dataset. *arXiv preprint arXiv:1611.09268* (2018).

[2] Keping Bi, Qingyao Ai, and W Bruce Croft. 2020. A transformer-based embedding model for personalized product search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1521–1524.

[3] Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2021. mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset. *CoRR* abs/2108.13897 (2021). arXiv:2108.13897 https://arxiv.org/abs/2108.13897

[4] Jason Ingyu Choi, Surya Kallumadi, Bhaskar Mitra, Eugene Agichtein, and Faizan Javed. 2020. Semantic product search for matching structured product catalogs in e-commerce. *arXiv preprint arXiv:2008.08180* (2020).

[5] Qiao Jin, Chuanqi Tan, Mosha Chen, Ming Yan, Songfang Huang, Ningyu Zhang, and Xiaozhong Liu. 2020. Aliababa DAMO Academy at TREC Precision Medicine 2020: State-of-the-art Evidence Retriever for Precision Medicine with Expert-in-the-loop Active Learning. In *TREC*.

[6] Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2022. COLIEE 2022 Summary: Methods for Legal Document Retrieval and Entailment. *Proceedings of the Sixteenth International Workshop on Juris-informatics (JURISIN 2022)* (2022).

[7] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies* 14, 4 (2021), 1–325.

[8] Yixiao Ma, Yunqiu Shao, Bulou Liu, Yiqun Liu, Min Zhang, and Shaoping Ma. 2021. Retrieving legal cases from a large-scale candidate corpus. *Proceedings of the Eighth International Competition on Legal Information Extraction/Entailment, COLIEE2021* (2021).

[9] Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2020. SLEDGE-Z: A Zero-Shot Baseline for COVID-19 Literature Search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4171–4179.

[10] Rodrigo Nogueira, Zhiying Jiang, Kyunghyun Cho, and Jimmy Lin. 2020. Navigation-based candidate expansion and pretrained language models for citation recommendation. *Scientometrics* 125, 3 (2020), 3001–3016.

[11] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 708–718.

[12] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2066–2070.

[13] Ronak Pradeep, Xueguang Ma, Xinyu Zhang, Hang Cui, Ruizhou Xu, Rodrigo Nogueira, and Jimmy Lin. 2020. H2oloo at trec 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine. *Corpus* 5, d3 (2020), d2.

[14] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2021. Summary of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. *Proceedings of the Eighth International Competition on Legal Information Extraction/Entailment* (2021).

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[16] Kirk Roberts, Dina Demner-Fushman, E. Voorhees, W. Hersh, Steven Bedrick, Alexander J. Lazar, and S. Pant. 2019. Overview of the TREC 2019 Precision Medicine Track. *The ... text REtrieval conference : TREC. Text REtrieval Conference* 26 (2019).

[17] Guilherme Moraes Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. No Parameter Left Behind: How Distillation and Model Size Affect Zero-Shot Retrieval. *arXiv preprint arXiv:2206.02873* (2022).

[18] Guilherme Moraes Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Roberto Lotufo, and Rodrigo Nogueira. 2022. Billions of Parameters Are Worth More Than In-domain Training Data: A case study in the Legal Case Entailment Task. *arXiv preprint arXiv:2205.15172* (2022).

[19] Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. To Tune or Not To Tune? Zero-shot Models for Legal Case Entailment. *ICAIL'21, Eighteenth International Conference on Artificial Intelligence and Law, June 21–25, 2021, São Paulo, Brazil* (2021).

[20] 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 preprint arXiv:2104.08663* (4 2021). https://arxiv.org/abs/2104.08663

[21] Ellen M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, November 16-19, 2004* (2004).

[22] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. (2021). arXiv:cs.CL/2010.11934

[23] Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020.*