Solution of Team GraphMIRAcles in the KDD Cup 2022 Query-Product Ranking Task

Hanzhu Chen, Zhihao Shi, Zhanqiu Zhang, and Jie Wang*
University of Science and Technology of China
Contact: jiewangx@ustc.edu.cn

ABSTRACT

To improve the users' shopping online experience, a search engine aims to show a ranked list of items that best match a user's query intent. The Query-Product Ranking task is formulated as search relevance to rank a given query-item pair by relevant labels: exact, substitute, complement, or irrelevant (ESCI) in the Shopping Queries Dataset, a large dataset of difficult Amazon search queries and results. However, many existing pre-trained models suffer from several challenges: noise in the data, inadaptation to the product data, slow convergence, and overfitting during fine-tuning. To address these challenges, we use the following methods in three components—data preprocessing, pre-training, and fine-tuning, respectively. We use regular expressions to clean the data and preprocess the data through data splicing, keyword extraction, and key sentence extraction. Then, we adapt our model to the domain corpus by Masked Language Model (MLM) pre-training. Finally, we use ranking loss in fine-tuning to accelerate convergence. To reduce model overfitting and improve model robustness, we use Fast Gradient Method (FGM) adversarial training.

Experiments demonstrate that our solution achieves an nDCG of 0.9002 on the private test dataset with a single model and can rank among the top 10 teams. By using ensemble methods, our models achieve an nDCG of 0.9028 on private test data and came fourth on the leaderboard.

Basic model	Model settings	nDCG (valid)
xlm-Roberta-large	Re	0.8958
xlm-Roberta-large	Dp + Re	0.8963
xlm-Roberta-large	Dp + Pt + Re	0.8993
xlm-Roberta-large	Dp + Pt + Ra	0.9012
xlm-Roberta-large	Dp + Pt + Ra + FGM	**0.9018**

Table 1. Experiment results of different model settings. Re: Regression loss; Dp: Data preprocessing; Pt: MLM pre-training; Ra: Ranking loss; FGM: FGM adversarial training.

* Corresponding author.

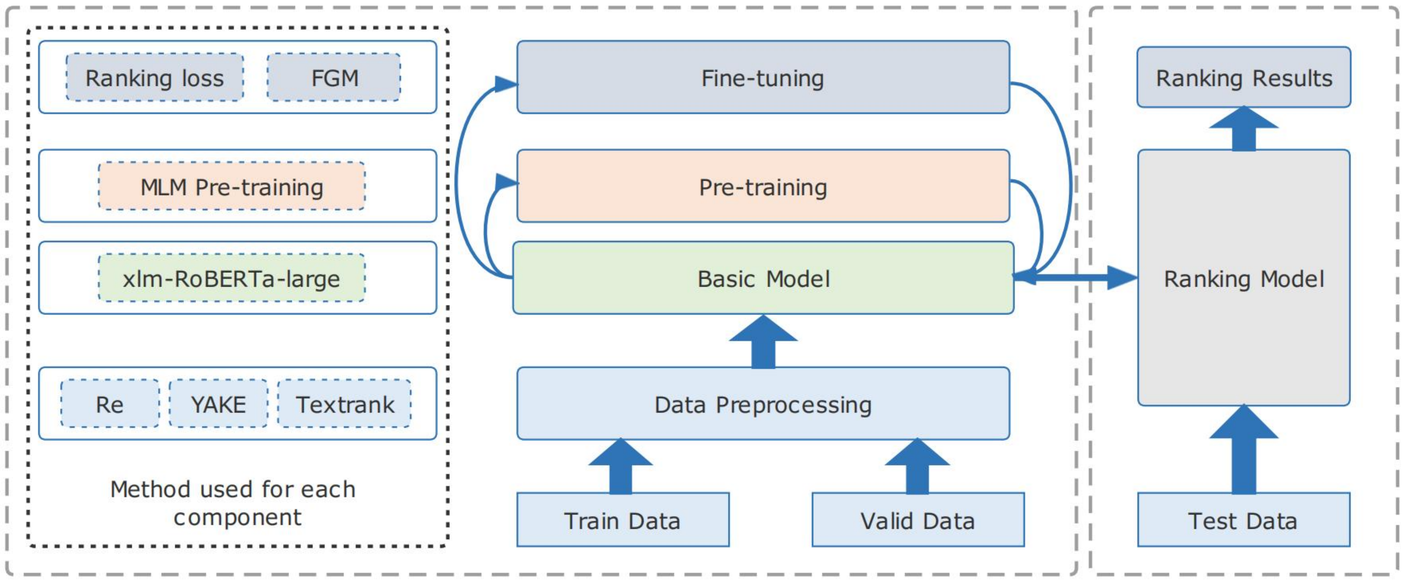

Figure 1. The overall architecture of our method. In the middle of the figure is a schematic diagram of each component. The left is the method used in each component, and their positions and colors correspond to the comp

Model	nDCG (private test)
Single Model	0.9002
Ensemble Model	**0.9028**

Table 2. Experiment results of the single model and the ensemble model

INTRODUCTION

Our solution consists of three components—data preprocessing, pre-training, and fine-tuning. Our solution is based on the basic model xlm-Roberta [1]. In the data preprocessing part, we use regular expressions to clean the data. In addition, we use YAKE [2] algorithm for keyword extraction and Textrank [3] algorithm for key sentence extraction. In the pre-training part, we allow the model to continue MLM [4] pre-training to adapt to the domain corpus. In the fine-tuning part, we use ranking loss to accelerate convergence and use FGM [5] adversarial training to reduce model overfitting and improve the robustness of the model. The overall architecture of our method is shown in Figure 1.

RESULTS

The results for different model settings are shown in Table 1. Briefly speaking, the xlm-Roberta-large model containing all components achieves the best nDCG results. Furthermore, Table 1 illustrates the effectiveness of our components.

References

[1] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guil laume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019).

[2] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. Information Sciences 509 (2020), 257–289.

[3] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.404–411.

[4] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964 (2020).

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).