Cross-Lingual Training of Dense Retrievers for Document Retrieval

Peng Shi[1], **Rui Zhang**[2], **He Bai**[1], **and Jimmy Lin**[1]

[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] Department of Computer Science and Engineering, Penn State University

{peng.shi,he.bai,jimmylin}@uwaterloo.ca, rmz5227@psu.edu

Abstract

Dense retrieval has shown great success for passage ranking in English. However, its effectiveness for non-English languages remains unexplored due to limitation in training resources. In this work, we explore different transfer techniques for document ranking from English annotations to non-English languages. Our experiments reveal that zero-shot model-based transfer using mBERT improves search quality. We find that weakly-supervised target language transfer is competitive compared to generation-based target language transfer, which requires translation models.

1 Introduction

Dense retrieval uses dense vector representations for semantic encoding and matching. However, most existing work focuses on high-resource languages such as English, where large-scale annotations are readily accessible. In this work, we focus on improving term-matching retrieval for low(er)-resource languages. We explore techniques for leveraging relevance judgments in English to train dense retrievers for document retrieval in non-English languages. Our experimental results show that combining dense retrieval and term-matching retrieval can obtain effectiveness improvements. Also, weakly-supervised target language transfer yields effectiveness competitive to generation-based target language transfer. This extended abstracted is an abridged version of Shi et al. (2021).

2 Cross-Lingual Relevance Transfer

We leverage the DPR model of Karpukhin et al. (2020), but using mBERT as the backbone model. During inference, we apply both bag-of-words exact term matching and dense retrieval. The relevance score of each document combines term-matching scores with dense retrieval similarity via $S_{doc} = \alpha \cdot S_{term} + (1 - \alpha) \cdot S_{dense}$, where α is tuned via cross-validation.

Model-Based Transfer. By exploiting the zero-shot cross-lingual transfer ability of pretrained transformers such as mBERT (Devlin et al., 2019), we train the dense retriever in the source language and apply inference directly to the target languages.

Target Language Transfer. To bridge the language gap between training and inference, we explore two techniques for creating a target language transfer set. (1) *Generation-based query synthesis*, where the goal is to leverage powerful generation models to predict reasonable queries given documents in the target language. We choose mBART (Liu et al., 2020) as our query generation model. The input of the model is the passage and its learning target is the corresponding query. We use the translate–train technique to obtain the generation models. More specifically, we leverage Google Translate to translate English query–document pairs into the target languages. Then, we use passages in the target language collections as input and generate corresponding queries in the same language. (2) *Weakly-supervised query synthesis.* We can automatically build the target language transfer set without manual annotation effort by treating the titles of Wikipedia articles as queries and the corresponding documents as positive candidates. We also retrieve top 1000 documents with BM25 for each query; documents except for the positive candidate are labeled as negative.

Two-Stage Training. We apply two-stage training to learn the dense retrieval model. The encoders are first trained on source language (English) annotated data which are available in larger quantities; then the models are fine-tuned on the synthesized query–document pairs in the target language.

3 Experimental Setup

We conduct experiments on six test collections: NTCIR 8 in Chinese, TREC 2002 in Arabic, CLEF 2006 in French, FIRE 2012 in Hindi, FIRE 2012

	AP	P@20	nDCG	AP	P@20	nDCG	AP	P@20	nDCG
Model		**NTCIR8-zh**			**TREC2002-ar**			**CLEF2006-fr**	
(**0**) BM25	0.4014	0.3849	0.4757	0.2932	0.3610	0.4056	0.3111	0.3184	0.4458
(**1**) BM25+RM3	0.3384	0.3616	0.4490	0.2783	0.3490	0.3969	0.3421	0.3408	0.4658
(**2**) NQ zero-shot	0.4221▲	0.4164▲	0.5235▲	0.2943	0.3560	0.4012	0.3470	0.3469	0.4726
(**3**) MS zero-shot	0.4167▲	0.4164▲	0.5095▲	0.3024	0.3810▲	0.4285	0.3332	0.3418	0.4573
(**4**) MS → QGen	0.4258▲	0.4336▲	0.5308▲	0.2988	0.3800	0.4276	0.3331	0.3429	0.4564
(**5**) MS → Wiki	0.4135	0.4123▲	0.5055▲	0.3060▲	0.3750	0.4293	0.3456	0.3480	0.4743
		FIRE2012-hi			**FIRE2012-bn**			**TREC3-es**	
(**0**) BM25	0.3867	0.4470	0.5310	0.2881	0.3740	0.4261	0.4197	0.6660	0.6851
(**1**) +RM3	0.3660	0.4430	0.5277	0.2833	0.3830	0.4351	0.4912	0.7040	0.7079
(**2**) NQ zero-shot	0.3939	0.4560	0.5408	0.2898	0.3980	0.4495▲	0.4910	0.6980	0.7007
(**3**) MS zero-shot	0.3944	0.4580	0.5461	0.2896▲	0.3900	0.4449	0.4950	0.7080	0.7171
(**4**) MS → QGen	0.3941	0.4660	0.5527	0.2887	0.3980	0.4486	0.4958▲	0.7180	0.7239
(**5**) MS → Wiki	0.3950	0.4630	0.5497	0.2898▲	0.4050	0.4549	0.4972▲	0.7180	0.7329

Table 1: Experimental results on baselines and our cross-lingual transfer methods. Model (0) and Model (1) show the effectiveness of BM25 and BM25 with RM3 query expansion. For each language, we select the higher P@20 of the two models as the term-based matching baseline. That is, for the French, Bengali, and Spanish collections, we use BM25+RM3 as the term-based matching baseline and for the others, we use BM25. Significant gains against the baselines are denoted with ▲.

in Bengali, TREC 3 in Spanish. For the evaluation metrics, we adopt AP, P@20, and nDCG@20. For model-based transfer, we explore Natural Question and MS MARCO as training datasets. For training the query generator in the target languages, we obtain training data by sampling 2000 query–passage pairs from MS MARCO and translate them into the target languages. Fisher's two-sided, paired randomization test (Smucker et al., 2007) at $p < 0.05$ was applied to test for statistical significance.

4 Results and Discussions

Finding #1: Zero-shot model-based transfer improves term-based matching. The results of zero-shot model-based transfer are shown in Model (2) and Model (3). Comparing with the corresponding baselines, we observe that model-based transfer, either NQ zero-shot or MS zero-shot, can improve retrieval effectiveness on P@20 for all collections, except NQ zero-shot on the TREC3-es dataset. We do not observe a clear winner between NQ and MS, though. These results indicate that mBERT-based DPR effectively transfers relevance matching across languages.

Finding #2: Target language transfer benefits certain collections, and Wiki query synthesis is better than query generation. Target language transfer results are shown in Model (4) and Model (5). MS → QGen and MS → Wiki denote the two-stage training strategy with different transfer sets, where QGen denotes generation-based query synthesis and Wiki denotes weakly-supervised query synthesis from Wikipedia. By comparing Model (4) with Model (3), we observe that second-stage training with generation-based query–document pairs can improve the effectiveness of P@20 over zero-shot model-based transfer on the Chinese, French, Hindi, Bengali, and Spanish collections. However, we see little improvement in terms of AP for all collections. By comparing Model (5) with Model (3), we find that second-stage training with weakly-supervised training data can improve P@20 over the zero-shot baselines on French, Hindi, Bengali, and Spanish.

Furthermore, by comparing these two transfer sets, we observe that, except for Chinese, Wiki obtains better retrieval effectiveness than QGen, which requires translation models for the target languages (which are expensive to build and not available for all languages).

5 Conclusion

We investigate the effectiveness of three transfer techniques for document ranking from English training data to non-English target languages. Our experiments in six languages demonstrate that zero-shot transfer using mBERT-based dense retrieval models improves term-based matching methods, and fine-tuning on augmented data in target languages can further benefit certain collections.

References

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.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick 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)*, pages 6769–6781, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-lingual training with dense retrieval for document retrieval. *arXiv preprint arXiv:2109.01628*.

Mark D Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632.