

CFDA & CLIP at TREC 2022 NeuCLIR Track

Jia-Huei Ju¹, Wei-Chih Chen¹, Heng-Ta Chang¹, Cheng-Wei Lin¹,
Ming-Feng Tsai², and Chuan-Ju Wang¹

¹Research Center for Information Technology Innovation, Academia Sinica

²Department of Computer Science, National Chengchi University

Abstract

In this notebook paper, we report our methods for NeuCLIR track in TREC 2022. We adopt the common multistage pipeline for cross-language information retrieval task (CLIR). The pipeline includes the machine translation, sparse passage retrieval and the cross-language passage re-ranking. Particularly, we fine-tune the cross-language passage re-rankers with different setting of query formulation. In the HC4 evaluation, our passage re-rankers achieved better passage re-ranking effectiveness compared to the baseline multilingual rerankers. The evaluation results on HC4 and NeuCLIR are also reported.

1 Methods

Our multistage pipeline is comprised of three stages: (i) machine translation, (ii) sparse passage retrieval and (iii) cross-language passage Re-ranking.

1.1 The Multi-stage Pipeline

Machine Translation. Before retrieval, we first translate the query q from the source language (i.e., English) to the machine-generated query \hat{q}_l , in target language l . We have tried different set of machine translation methods, including mT5 [6], NLLB [5]¹. However, we found that the officially provided translation via Google translation API shows the highest recall among these variants. We use Google’s translated query for the further retrieval and re-ranking.

Sparse Retrieval. With the machine-translated queries, we regard the CLIR task as a monolingual ad-hoc passage retrieval task. For each language l , we retrieve the top-1000 relevant documents \bar{D}_l via BM25 search as

$$\bar{D}_l = \phi_{\text{BM25}}(\hat{q}_l, \bar{D}_l),$$

where ϕ_{BM25} indicates the sparse passage retrieval model. Additionally, following [4], we adopted query expansion approach (RM3) built in Pyserini [3] to increase the recall in this stage.

Cross-language Passage Re-ranking We further re-rank the retrieved candidate passages using cross-encoder models. Particularly, we aim to leverage the semantic meaning in the original query q as well as the translated query q_l , as we hypothesize the underlying information loss in machine translation may negatively affect the effectiveness. Our passage re-rankers are fine-tuned with mT5 models [6], and can be formulated as

$$R = \phi_{\text{mT5}}(q; q_l, d_l),$$

where the ϕ_{mT5} is one of our re-rankers (See detail in 1.2). d_l represents the documents retrieved from the last stage. R is the final ranked list of the results.

¹<https://github.com/facebookresearch/fairseq/tree/nllb>

1.2 Fine-tuning Cross-language Passage Ranking

In this section, we introduce passage re-rankers with other two different settings. Follow the baseline mT5 reranker [1], the text-to-text formulation of query-passage pair is as follow,

Query: q_l Document: d_l Relevant:.

Moreover, we recast the text-to-text formulations with two different query settings:

1. Bilingual query: in addition to the monolingual query-passage pair (q, p_l) , we concatenate the original English query q in the beginning.

Query: q Query Translation: q_l Document: d_l Relevant:.

2. crosslingual query: such setting is similar to end-to-end CLIR objectives [4] and is formulated with query and passage in different languages.

Query: q Document: d_l Relevant:.

With these formulations, we can thereby leverage the training triplet in mMARCO [1] (i.e, the translated MSMARCO in 14 languages). The triplet data includes query and positive/negative passages; can thereby formulated as the **yes/no** tokens generation tasks [1]. All the settings are identical to the original mT5 passage re-ranker, such as batch size, training steps, learning rate, ...etc.

2 Experiments

In the experiments, we validate the effectiveness of proposed cross-language passage re-ranking. We report the settings with empirical results on HC4 testing set.

2.1 Settings

Dataset. As for training data, we use the mMARCO dataset [1] to construct the training examples. Unlike the original mT5 re-ranking model [1] using 9 languages, we only use Russian, Chinese and English as our fine-tuning triplets because we think the fewer languages used can help us analyzing the experiment results more clear. As for evaluation, we use CLIR Common Crawl Collection (HC4) [2] as our testing data; this testing data have 50, 50, 62 queries respectively, in Chinese, Persian and Russian languages.

Compared Rerankers. We regard two re-rankers baselines in our experiments, including

1. **mT5-orig**, the original baseline reranker fine-tune on 9 languages with randomly distributed *monolingual* triplets. We directly use the fine-tuned checkpoints in our experiments.².
2. **mT5-mono**, we randomly distributed *monolingual* triples among three languages (English, Chinese and Russian) since there are only two target languages (Russian and Chinese) matched in mMARCO and this track.

As mentioned in Section 1.2, we use the other two cross-language query settings to fine-tune the cross-language passage re-rankers. For both of the settings, we randomly select Chinese or Russian as the target language for cross-language query q_l ; and further formulate the source input accordingly. The mT5 rerankers fine-tuned with bilingual query and crosslingual query are named as **mT5-bi** and **mT5-cl**.

Fixed First-stage Retrieval. To alleviate the information loss of translation, we adopt the first-stage retrieval with **human** translated query. All the retrieval in each three languages are performed by BM25 search with top 1000 passages.

2.2 Re-ranking results on HC4

We report the empirical results of the rerankers on HC4 testing dataset in the Table 2 with Recall, nDCG, MAP at different cut-off. We separate our results into three blocks in different target languages (e.g. Persian, Russian and Chinese). The numbers in boldface indicate the highest among our compared methods.

²<https://huggingface.co/unicamp-dl/mt5-base-mmarco-v2>

Table 1: The results of passage re-ranking on HC4 testing set, including persian, chinese and russian.

Rerankers	Size	R@100	nDCG@20	mAP@20	MAP@100	MAP@1K
Persian (fas)						
mT5-orig	base	0.7175	0.4726	0.3311	0.3626	0.3666
mT5-mono	large	0.7602	0.5488	0.3987	0.4253	0.4285
mT5-bi	large	0.7600	0.5644	0.4123	0.4411	0.4442
mT5-cl	large	0.7648	0.5491	0.4078	0.4296	0.4330
Russian (rus)						
mT5-orig	base	0.5923	0.2946	0.2016	0.2512	0.2599
mT5-mono	large	0.6752	0.3698	0.2564	0.3168	0.3243
mT5-bi	large	0.6860	0.3822	0.2768	0.3377	0.3450
mT5-cl	large	0.6595	0.3757	0.2603	0.3172	0.3251
Chinese (zho)						
mT5-orig	base	0.7374	0.4928	0.3574	0.3949	0.4004
mT5-mono	large	0.7826	0.5778	0.4473	0.4817	0.4851
mT5-bi	large	0.7623	0.5743	0.4246	0.4574	0.4621
mT5-cl	large	0.7838	0.5924	0.4450	0.4794	0.4823

Table 2: The results of the full ranking results on NeuCLIR evaluation set

Rerankers	Runs	nDCG@20	mAP@20	MAP@100	MAP@1K
Persian — run prefix: CFDP_CLIP_fas_					
mT5-mono	(L)	0.4876	0.2644	0.3178	0.3435
mT5-bi	(dq)	0.5077	0.2797	0.3398	0.3639
mT5-cl	(clf)	0.4681	0.2480	0.3034	0.3298
Russian — run prefix: CFDP_CLIP_rus_					
mT5-mono	(L)	0.4693	0.2212	0.3102	0.3486
mT5-bi	(dq)	0.5126	0.2553	0.3480	0.3862
mT5-cl	(clf)	0.5071	0.2534	0.3465	0.3829
Chinese — run prefix: CFDP_CLIP_zho_					
mT5-mono	(L)	0.4808	0.2402	0.3157	0.3454
mT5-bi	(dq)	0.4838	0.2570	0.3293	0.3603
mT5-cl	(clf)	0.4790	0.2448	0.3187	0.3494

Zero-shot Effectiveness (Persian Query). In the first block in Table 2, we can regard the Persian (fas) language settings as the *zero-shot* CLIR task since the Persian text is not included in our fine-tuning triples. We observe that both of our proposed cross-language passage re-rankers (mT5-bi, mT5-cl) outperform the baselines (mT5-orig and mT5-mono) in the shallower depth, which imply that the cross-language query (e.g. bilingual query or crosslingual query) can potentially guide the representation of query-passage pair in different languages.

Effectiveness of Cross-language Query. As for the cross-language effectiveness, we compare our results in the last two blocks (i.e., Russian and Chinese). For the Russian CLIR task, we observe the *bilingual query* setting (mT5-bi) outperform the other rerankers. However, for the Chinese CLIR task, we observe that there are only minor improvement of our proposed approaches compared to the baseline mT5-mono. We hypothesize the inconsistent improvements between languages are derived from the inherent linguistic gap of different languages. Particularly, we find that the mT5-bi perform totally opposite in Russian and Chinese (the highest in Russian, yet lowest in Chinese). As far as our understanding, the English-Chinese gap is inherently larger than English-Russian gap; therefore, the poor performance when we fine-tune our reranker with *bilingual query* when the linguistic gap between source-target language is larger.

2.3 Results on NeuCLIR

We also report the evaluation results of our submitted runs in NeuCLIR. Interestingly, we can observe that the *bilingual query* (i.e., run name with dq) are the best among all of the other setting. Although these results are not consistent to our development set (HC4 testing set), we can still conclude that the bilingual query setting is regarded as a promising approach for passage re-ranking. Furthermore, this result also show that encoding texts in multiple languages within the same input can bring more signals.

3 Conclusion

We evaluate the effectiveness of cross-language query-passage pairs, and aim to explore the better practice of fine-tuning text ranking models for ad-hoc cross-language information retrieval. In our empirical evaluation, we suggest that it is easier to achieve the decent performance when the source-target language gap is smaller in CLIR. As our future work, we will conduct more comprehensive evaluation on different languages and different benchmark dataset to learn more about the gaps of different languages. Also, we will further apply our settings on the self-supervised pre-training tasks for CLIR, which we aim to explore the effective multi-lingual pre-trained language models beyond the one training in monolingual manner.

References

- [1] L. Bonifacio, V. Jeronymo, H. Q. Abonizio, I. Campiotti, M. Fadaee, R. Lotufo, and R. Nogueira. mmarco: A multilingual version of the ms marco passage ranking dataset, 2021. URL <https://arxiv.org/abs/2108.13897>.
- [2] D. Lawrie, J. Mayfield, D. W. Oard, and E. Yang. Hc4: A new suite of test collections for ad hoc clir. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 351–366, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-030-99735-9. doi: 10.1007/978-3-030-99736-6_24. URL https://doi.org/10.1007/978-3-030-99736-6_24.
- [3] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, page 2356–2362, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463238. URL <https://doi.org/10.1145/3404835.3463238>.
- [4] S. Nair, E. Yang, D. Lawrie, K. Duh, P. McNamee, K. Murray, J. Mayfield, and D. W. Oard. Transfer learning approaches for building cross-language dense retrieval models, 2022. URL <https://arxiv.org/abs/2201.08471>.
- [5] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. Meja-Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. No language left behind: Scaling human-centered machine translation. 2022.
- [6] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL <https://aclanthology.org/2021.naacl-main.41>.