CFDA & CLIP at TREC 2022 NeuCLIR Track

Jia-Huei Ju[1], Wei-Chih Chen[1], Heng-Ta Chang[1], Cheng-Wei Lin[1],
Ming-Feng Tsai[2], and Chuan-Ju Wang[1]

[1]Research Center for Information Technology Innovation, Academia Sinica
[2]Department of Computer Science, National Chengchi University

Abstract

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

1 Methods

Our multi-stage pipeline is comprised of three stages: (i) machine translation, (ii) sparse passage retrieval and (ii) 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 the target language l. We have tried different sets of machine translation methods, including mT5 [6] and NLLB [5].[1] However, we found that the officially provided translation via the Google Translation API shows the highest recall among these variants. Finally, we use the Google-translated query for passage candidate retrieval and passage re-ranking.

Sparse Retrieval. With the machine-translated queries, we can 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

$$\bar{D}_l = \phi_{\text{BM25}}(\hat{q}_l, \bar{D}_l), \tag{1}$$

where ϕ_{BM25} indicates the sparse passage retrieval model. Additionally, following [4], we use sparse retrieval with pseudo relevance feedback technique; we use the 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 for more effective passage re-ranking, as we hypothesize that 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

$$R = \phi_{\text{mT5}}(q; q_l, d_l \in \bar{D}_l), \tag{2}$$

where ϕ_{mT5} is one of our re-rankers (see details in Section 1.2). d_l represents the documents retrieved from the last stage, and R is a final ranked list with re-ordered passages $d_l \in \bar{D}_l$.

[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 different settings that can achieve cross-language passage ranking. Following the baseline mT5 re-ranker [1], the text-to-text formulation is:

$$\texttt{Query: } q_l \texttt{ Document: } d_l \texttt{ Relevant:}.$$

where the query and document are in the same language l as described in Eq. 2. 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_l, d_l), we concatenate the original English query q in the beginning.

$$\texttt{Query: } q \texttt{ Query Translation: } q_l \texttt{ Document: } d_l \texttt{ Relevant:}$$

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

$$\texttt{Query: } q \texttt{ Document: } d_l \texttt{ Relevant:}.$$

We further use the training triplets in mMARCO [1] (i.e., the translated MSMARCO in 14 languages). The triplet data includes query and positive/negative passages; it can thereby be formulated as the `yes/no` tokens generation tasks [1]. All the other 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 focus on passage re-ranking and 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] which uses 9 languages, we only use Russian, Chinese, and English as our fine-tuning triplets because we think that using fewer languages can help us analyze the experiment results more clearly. As for evaluation, we use the CLIR Common Crawl Collection (HC4) [2] as our testing data; this testing data has 50, 50, and 62 queries respectively, in Chinese, Persian, and Russian languages.

Sparse Retrieval with Machine Translation. Before using BM25 search, we use Google Translate (G-Trans) as our first-stage machine translation.[2] With the original English query q and G-Trans query $\hat{q}_l$, we retrieve the top 1000 relevance passage candidates and pass them to the compared re-rankers.

Compared Re-rankers. We regard two re-ranker baselines in our experiments, including:

1. **mT5-orig**, the original baseline re-ranker fine-tuned on nine languages with randomly distributed monolingual triplets. We directly use the fine-tuned checkpoints in our experiments[3].
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. Particularly, for both of the settings, we randomly select Chinese or Russian as the target language l with the cross-language query q_l and formulate the source input instead of a monolingual manner (see Section 1.2). The mT5 re-rankers fine-tuned with bilingual query and cross-lingual query are named **mT5-bi** and **mT5-cl**, respectively.

[2] Note that we use the **human-translated** query on HC4 evaluation; while we use the **G-trans** query in our submitted runs for this track. We hypothesize the human-translated query can alleviate the information loss of translation, making it easier for us to analyze the effectiveness of our proposed re-rankers.

[3] `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. For re-rankers other than mT5-cl, we use the **human translated** query for BM25 search and re-rankers.

re-rankers	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

2.2 Re-ranking results on HC4

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

Zero-shot Effectiveness (Persian Query). In the first block of Table 2, we can regard the Persian (fas) language settings as a *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) at shallower depths, which implies that the cross-language query (e.g. bilingual query or crosslingual query) can potentially guide the representation of query-passage pairs 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 that the *bilingual query* setting (mT5-bi) outperforms the other re-rankers. However, for the Chinese CLIR task, we observe only minor improvements of our proposed approaches compared to the baseline mT5-mono. We hypothesize that the inconsistent improvements between languages are derived from the inherent linguistic gap of different languages. Particularly, we find that mT5-bi performs 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 the English-Russian gap; therefore, the performance is relatively poor when we fine-tune our re-ranker with *bilingual query* (i.e., mT5-bi), implying the linguistic gap between the source and 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) is the best among all of the other settings. Although these results are not consistent with the evaluation on HC4 testing set, we can still conclude that the bilingual query setting is a promising approach for passage re-ranking. This result also shows that encoding query in multiple languages contextually as a single input can bring signals for passage re-rankers.

Table 2: The results of the full ranking results on NeuCLIR evaluation set. For the re-rankers involved translated query, we use the **Google translated** query for BM25 search and re-rankers.

re-rankers	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

3 Conclusion

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