UMass at TREC 2023 NeuCLIR Track

Zhiqi Huang, Puxuan Yu, and James Allan

Center for Intelligent Information Retrieval
Manning College of Information and Computer Sciences
University of Massachusetts Amherst
{zhiqihuang,pxyu,allan}@cs.umass.edu

Abstract. This overviews the University of Massachusetts's efforts in cross-lingual retrieval run submissions for the TREC 2023 NeuCLIR Track. In this cross-lingual information retrieval (CLIR) task, the search queries are written in English, and three target collections are in Chinese, Persian, and Russian. We focus on building strong ensembles of initial ranking models, including dense and sparse retrievers.

1 Introduction

The NeuCLIR track is dedicated to advancing cross-lingual information retrieval (CLIR) studies by leveraging deep learning and neural models. NeuCLIR queries are articulated in English, while the document collections are composed in Chinese, Persian, and Russian. This year, besides three independent CLIR tasks, NeuCLIR also introduces two new tasks: (i) a multilingual retrieval task where the collections across all three languages are combined into a single corpus. (ii) a domain-specific retrieval task where the collection is Chinese technical documents (specifically, abstracts of academic papers and theses).

In this paper, we detail our TREC 2023 NeuCLIR track submissions. Broadly speaking, our approaches can be categorized into two types based on whether the retrieval method uses the translated documents provided by the organizer. Leveraging machine-translated (MT) documents from the target language to English shifts the retrieval challenge from cross-lingual to monolingual. Our submission in the monolingual retrieval setting is the fusion of multiple neural approaches and the NeuCLIR baseline. Another category is retaining the documents in their original languages and conducting retrieval with English queries. Because this year's NeuCLIR also has a multilingual retrieval task, instead of building separate models for each language pair, we build multilingual retrieval models and fuse the rank lists as the submission for cross-lingual setting. Finally, we combine the runs from both monolingual and cross-lingual settings to achieve a more comprehensive rank list. Our initial analysis of the evaluation results shows that retrieval models using MT documents outperform the CLIR models, indicating a high translation quality of three target languages to English.

2 Methodology

2.1 Translated Documents

The retrieval task transforms from cross-lingual to monolingual by employing the document translation supplied by the NeuCLIR organizer. We apply both dense and sparse neural retrieval models to search the English collections translated from target languages. For dense passage retriever, we leverage two publicly available models, coCondenser [Gao and Callan, 2021] and CoT-MAE [Wu et al., 2023], from the top of MS MARCO passage ranking leaderboard. For sparse retrieval, we leverage a pretrained SPLADE++ [Formal et al., 2022] to index the collection for English-English retrieval. Last year's (2022) NeuCLIR reported the baseline method as BM25 with document translation [Lawrie et al., 2023]. Therefore, we still consider the baseline from this year as a method using translated documents. We leverage Reciprocal Rank Fusion (RRF) to combine multiple retrieval results based on smoothed reciprocal rank. Our submission in monolingual strategy is a fusion of NeuCLIR baseline, coCondenser, CoT-MAE and SPLADE.

2.2 Native Documents

Rather than adopting the translate-then-retrieve approach, the cross-lingual challenge can also be tackled using CLIR models. We apply two modeling strategies: (i) fine-tuning existing models with cross-lingual retrieval data between English and target languages. (ii) building a new document encoder in the target languages by transferring the retrieval knowledge from English to the target languages. Our submission using native documents is also a fusion of multiple runs.

Training on CLIR Data: We leverage the translated MS MARCO passage collection to build CLIR models. We take Chinese and Russian translations from mMARCO [Bonifacio et al., 2021]. For Farsi, we use a collection translated by NeuCLIR 2022. We train both dense and sparse retrievers using CLIR data. First, we build optimized mDPR by fine-tuning mDPR on three target languages. We also finetune a single mSPLADE (initializing SPLADE with mDPR checkpoint) on the combined mMARCO and NeuCLIR collection, use the trained checkpoints to index this year's collections, and perform sparse retrieval. Note that we do not restrict the vocabulary to those only in the languages of interest as in BLADE [Nair et al., 2023].

Knowledge Transfer via Translation Data: Instead of directly learning from CLIR data, we can also build CLIR models by transferring existing English retrieval models to the target languages [Huang et al., 2023a]. Following the approach of KD-SPD [Huang et al., 2023b], we train a new multilingual document encoder by distilling ANCE [Xiong et al., 2020] using parallel sentences between English and three target languages. The original ANCE (query encoder) and the multilingual document encoder form a bi-encoder retrieval architecture.

Table 1: Results of our submission along with the averaged min/median/max of all submissions from NeuCLIR official. Task column indicating the language of the target collection. `mlir` refers to the multilingual collection.

Task	# Queries	Run	nDCG@20	Recall@1000
`fa`	60	min/median/max	0.1015 / 0.4400 / 0.7260	0.3338 / 0.8913 / 0.9875
		NativeFusion	0.4398	0.8834
		TransFusion	**0.5059**	**0.9580**
		Hybrid	0.4962	0.9356
`ru`	62	min/median/max	0.0645 / 0.4306 / 0.7002	0.3500 / 0.8518 / 0.9805
		NativeFusion	0.4284	0.8824
		TransFusion	**0.4999**	**0.9206**
		Hybrid	0.4976	0.9193
`zh`	62	min/median/max	0.0323 / 0.3864 / 0.7040	0.2469 / 0.8743 / 0.9872
		NativeFusion	0.3897	0.8924
		TransFusion	**0.4573**	**0.9365**
		Hybrid	0.4449	0.9393
`mlir`	65	min/median/max	0.1177 / 0.3723 / 0.5960	0.4909 / 0.8135 / 0.9200
		NativeFusion	0.3594	0.7799
		TransFusion	**0.4156**	**0.8559**
		Hybrid	0.4082	0.8453

Table 2: Results of domain-specific cross-lingual retrieval. `csl` represents the Chinese scientific literature dataset.

Task	# Queries	Run	nDCG@20	mAP
`csl`	39	min/median/max	0.0172 / 0.3272 / 0.6328	0.0105 / 0.2184 / 0.5069
		NativeFusion	0.2313	0.1479
		TransFusion	**0.3411**	**0.2354**
		Hybrid	0.3266	0.2203

2.3 Runs Submitted to TREC

For each target language and the multilingual collection, we submit three runs:

- TransFusion: The fusion of methods using translated documents. We combine rank lists from coCondenser, CoT-MAE, and SPLADE.

– NativeFusion: The fusion of methods using native documents. We combine rank lists from mDPR-optimized, KD-SPD, and mSPLADE.
– Hybrid: The fusion of NativeFusion and TransFusion.

All our submissions are (i) automatic runs, (ii) based on English queries, and (iii) search the complete document collection (first-stage retriever).

3 Initial Analysis on TREC NeuCLIR 2023

For each query, the NeuCLIR officially evaluates our submission and provides the minimum, median, and maximum evaluation metrics of all participants for comparison. Table 1 shows the results of our submissions compared with statistics from all NeuCLIR participants. We report the official nDCG score at the top 20 and recall at the top 1000 retrieved documents.

We make the following observations: First, the retrieval results using translated documents (TransFusion) surpass the performance in cross-lingual contexts (NativeFusion) across all languages, indicating a high translation quality from three target languages to English. Additionally, our NativeFusion performs close to the median of all submissions. And TransFusion significantly outperforms the median in terms of both nDCG and recall. Finally, the combination of NativeFusion and TransFusion (Hybrid) does not show an improvement. Despite the differences in language, models in NativeFusion have a similar model architecture and are learning the same retrieval knowledge in cross-lingual settings as models in TansFusion. The retrieval knowledge is learned from the label of MS MARCO passage ranking dataset. This limited the improvement when combining NativeFusion and TransFusion. To further improve the retrieval performance, more data, especially in-domain retrieval data, are required for model training.

Table 2 shows the results of domain-specific cross-lingual retrieval. The collection is Chinese academic papers and theses. For each paper, we concatenate the title, keywords, and abstract to form a document. We can see that in this domain, NativeFusion is below the median, indicating the CLIR methods are significantly affected by the domain shift.

4 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Bonifacio, L., Jeronymo, V., Abonizio, H.Q., Campiotti, I., Fadaee, M., Lotufo, R., Nogueira, R.: mmarco: A multilingual version of the ms marco passage ranking dataset. arXiv preprint arXiv:2108.13897 (2021)

Formal, T., Lassance, C., Piwowarski, B., Clinchant, S.: From distillation to hard negative sampling: Making sparse neural ir models more effective. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2353–2359 (2022)

Gao, L., Callan, J.: Unsupervised corpus aware language model pre-training for dense passage retrieval. arXiv preprint arXiv:2108.05540 (2021)

Huang, Z., Yu, P., Allan, J.: Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, pp. 1048–1056 (2023a)

Huang, Z., Zeng, H., Zamani, H., Allan, J.: Soft prompt decoding for multilingual dense retrieval. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 1208–1218, SIGIR '23 (2023b)

Lawrie, D., MacAvaney, S., Mayfield, J., McNamee, P., Oard, D.W., Soldaini, L., Yang, E.: Overview of the trec 2022 neuclir track. arXiv preprint arXiv:2304.12367 (2023)

Nair, S., Yang, E., Lawrie, D., Mayfield, J., Oard, D.W.: Blade: Combining vocabulary pruning and intermediate pretraining for scaleable neural clir (2023)

Wu, X., Ma, G., Lin, M., Lin, Z., Wang, Z., Hu, S.: Contextual masked autoencoder for dense passage retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 4738–4746 (2023)

Xiong, L., Xiong, C., Li, Y., Tang, K.F., Liu, J., Bennett, P., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. arXiv preprint arXiv:2007.00808 (2020)