HUAWEI NOAH'S ARK LAB AT TREC NEUCLIR 2022

Ehsan Kamalloo*
University of Alberta, Canada
kamalloo@ualberta.ca

David Alfonso-Hermelo
Huawei Noah's Ark Lab, Canada
david.alfonso.hermelo@huawei.com

Mehdi Rezagholizadeh
Huawei Noah's Ark Lab, Canada
mehdi.rezagholizadeh@huawei.com

ABSTRACT

In this paper, we describe our participation in the NeuCLIR track at TREC 2022. Our focus is to build strong ensembles of full-ranking models including dense retrievers, BM25 and learned sparse models.

1 Introduction

The NeuCLIR track, launched for the first time at TREC 2022, aims at fostering cross-lingual retrieval research where topics are specified in English and document collections are written in Chinese, Persian, and Russian. In this paper, we describe our submissions to the NeuCLIR 2022. Our main strategy is to test existing full-ranking retrieval models and build an ensemble model over them. We did not adopt any re-ranking methods this year.

2 Methodology

We follow the query translation paradigm [Nie, 2010] where English queries are translated into the language of documents often using an off-the-shelf translation tool. We build a bespoke model for each language, as opposed to employing one multilingual model for all languages. To this end, we first train a dense retrieval model based on the widely adopted two-tower architecture of *bi-encoders* [Lin et al., 2021a]. Then, we build an ensemble model by combining our dense retrievers with other strong baselines.

Dense Retrieval: Our retriever is an XLM-RoBERTa [Conneau et al., 2020], following [Nair et al., 2022]. We use Tevatron [Gao et al., 2022] to train a bi-encoder model with shared weights. We first fine-tune XLM-RoBERTa$_{\texttt{large}}$[2] (XLM-R in short) on English MS MARCO, as done in Zhang et al. [2022]. Then, we initialize our non-English retrievers by the weights of the English model and fine-tune them on their corresponding document collections.

Our baselines include BM25, provided in Pyserini [Lin et al., 2021b], and SPLADE, a prominent learned sparse retrieval model [Formal et al., 2021].

2.1 Ensembling

We leverage Reciprocal Rank Fusion (RRF) [Cormack et al., 2009] to combine various retrieval models. RRF is based on smoothed reciprocal rank and equally treats all retrieval models. Sometimes, one model should influence the output more than others because it has more relevant documents at top ranks. To account for this case, we adjust the original RRF by assigning a weight for each input model. More precisely, the RRF score for document d given a set of retrieval models R is modified as the following:

*Work done while at Huawei Noah's Ark Lab.

[2] https://huggingface.co/xlm-roberta-large

$$RRF_{\text{score}}(d) = \sum_{r \in R} \frac{\alpha_r}{k + r(d)} \tag{1}$$

where α_r denotes the corresponding weight of a retriever.

The retrieval model weights in our modified RRF are hyper-parameters. We tune them by conducting an exhaustive search over all possible combinations.

2.2 Query Translation

In addition to the provided human translations, we tested our models using five various translation tools including Caiyun[3] (ca), Huawei[4] (hw), Facebook NLLB[5] (fb) [Costa-jussà et al., 2022], Microsoft Bing Translator[6] (msr), and Youdao[7] (you). Note that not all target languages are supported by these systems, e.g., Caiyun does not offer translation for Persian.

3 Results

3.1 Datasets

The NeuCLIR track offers two document corpora, HC4 [Lawrie et al., 2022] and NeuCLIR-1 for development and test, respectively. Both corpora are collected from the Common Crawl news collection. To identify documents in Chinese, Persian, and Russian, the language of documents were determined via automated language identification. HC4 comprises documents within a three-year time frame, i.e., August 2016 to August 2019, whereas documents in NeuCLIR-1 are taken from August 2016 to July 2021.

For training a dense retriever, we leverage mMARCO [Bonifacio et al., 2021], i.e., an automatically translated version of MS MARCO [Bajaj et al., 2016] that accounts for 13 languages. However, mMARCO does not include Persian, which is why we translated MS MARCO into Persian using mBART-large[8] [Tang et al., 2020].

3.2 Dense Retrieval

The results of our dense retrievers on the HC4 test data are presented in Table 1. We observe that using "title+description" yields the best results across all the three languages. In this experiment, we used the provided human translations for queries.

Table 1: nDCG@20 of XLM-RoBERTa dense retrievers on the HC4 test data

Language	Query		
	title	desc	title+desc
`zh`	0.151	0.138	0.154
`fa`	0.173	0.207	0.232
`ru`	0.129	0.167	0.182

3.3 Ensemble Retrieval

For each language, we generated up to 15 runs to fuse with our dense retrievers based on the combination of query fields (title and description), translation tools (§2.2), and baselines (BM25 and SPLADE) with or without pseudo relevance feedback (PRF). We combined a dense retriever run with up to two other runs. The RRF weights are selected from $\{1, 2\}$ and are determined by enumerating all combinations to find the best configurations on the HC4 dev set.

We made three submissions for each language from the top-3 configurations that we found. For two submissions, represented by `c-hybrid2` and `c-hybrid3`, we excluded runs that use human translation from the tuning step to mimic

[3] `https://fanyi.caiyunapp.com/`

[4] An internal translation tool

[5] `https://huggingface.co/facebook/nllb-200-distilled-600M`

[6] `https://www.bing.com/translator`

[7] `https://fanyi.youdao.com/`

[8] `https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt`

the "true" cross-lingual retrieval and for the third submission, denoted by `m-hybrid1`, we considered all runs in the tuning phase. The details of our submissions are presented in Table 2.

Table 2: nDCG@20 and Recall@1k of our submissions on the HC4 test data along with the official results on the NeuCLIR topics. t=title, d=description, and ×2 indicates the model weight in RRF.

Lang.	**Run**	**Ensemble**	**HC4** nDCG@20	**HC4** R@1k	**NeuCLIR** nDCG@20	**NeuCLIR** R@1k
zh	m-hybrid1	XLM-R (*t*) {human} ×2 XLM-R (*d*+*t*) {you} BM25 (*t*,*d*+*t*) {human, ca, hw, msr, you} ×2	0.157	0.631	0.390	0.756
	c-hybrid2	XLM-R (*d*) {msr} ×2 XLM-R (*t*) {you} BM25 (*t*,*d*+*t*) {ca, hw, msr, you} ×2	0.153	0.619	0.359	0.703
	c-hybrid3	XLM-R (*d*) {ca} XLM-R (*t*) {you} BM25 (*t*,*d*+*t*) {human, ca, hw, msr, you}	0.149	0.623	0.372	0.706
fa	m-hybrid1	XLM-R (*d*) {fb} SPLADE+PRF (*t*+*d*) {hw} ×2 BM25+PRF (*t*,*d*+*t*) {human, fb, hw, msr} ×2	0.484	0.922	0.467	0.897
	c-hybrid2	XLM-R (*t*+*d*) {fb} SPLADE+PRF (*t*+*d*) {msr} ×2 SPLADE+PRF (*t*+*d*) {fb}	0.273	0.746	0.411	0.845
	c-hybrid3	XLM-R (*t*+*d*) {msr} SPLADE+PRF (*t*+*d*) {msr} ×2 SPLADE+PRF (*t*+*d*) {fb}	0.264	0.746	0.415	0.845
ru	m-hybrid1	XLM-R (*d*) {human} SPLADE (*t*+*d*) {hw} ×2 BM25+PRF (*t*,*d*+*t*) {human, hw, msr} ×2	0.245	0.781	0.501	0.881
	c-hybrid2	XLM-R (*d*) {msr} SPLADE (*t*+*d*) {hw} ×2 BM25+PRF (*t*,*d*+*t*) {hw, msr} ×2	0.239	0.784	0.496	0.879
	c-hybrid3	XLM-R (*d*) {hw} SPLADE (*t*+*d*) {msr} ×2 BM25+PRF (*t*,*d*+*t*) {hw, msr} ×2	0.245	0.764	0.493	0.877

References

Jian-Yun Nie. Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*, 3(1):1–125, 2010.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. Pretrained transformers for text ranking: BERT and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325, 2021a. doi:10.2200/S01123ED1V01Y202108HLT053.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.747.

Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. Transfer learning approaches for building cross-language dense retrieval models. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*, pages 382–396. Springer International Publishing, 2022. doi:10.1007/978-3-030-99736-6_26.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Tevatron: An efficient and flexible toolkit for dense retrieval. 2022. doi:10.48550/arXiv.2203.05765.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. Towards best practices for training multilingual dense retrieval models. 2022. doi:10.48550/arXiv.2204.02363.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo 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*, pages 2356–2362. Association for Computing Machinery, 2021b. doi:10.1145/3404835.3463238.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE v2: Sparse lexical and expansion model for information retrieval. 2021. doi:10.48550/arXiv.2109.10086.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759. Association for Computing Machinery, 2009. doi:10.1145/1571941.1572114.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. 2022. doi:10.48550/arXiv.2207.04672.

Dawn Lawrie, James Mayfield, Douglas W Oard, and Eugene Yang. HC4: A new suite of test collections for ad hoc CLIR. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*, pages 351–366. Springer International Publishing, 2022. doi:10.1007/978-3-030-99736-6_24.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mMARCO: A multilingual version of the ms marco passage ranking dataset. 2021. doi:10.48550/arXiv.2108.13897.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated MAchine Reading COmprehension dataset. 2016. doi:10.48550/arXiv.1611.09268.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. 2020. doi:10.48550/arXiv.2008.00401.