HNUST @ TREC 2022 NeuCLIR Track

Ge Zhang, Qiwen Ye, Mengmeng Wang and Dong Zhou (✉)

School of Computer Science and Engineering, Hunan University of Science and Technology, HNUST, Hunan 411201, China

Abstract. With the rapid development of deep learning, neural-based cross-language information retrieval (CLIR) has attracted extensive attention from researchers. To explore the effectiveness of neural-based CLIR, large-scale efforts, and new platforms are in need. To that end, the TREC 2022 NeuCLIR track presents a cross-language information retrieval challenge. This paper describes our first participation in the TREC 2022 NeuCLIR track. We explored two approaches for CLIR: (1) the lexical-based CLIR method and (2) the neural-based CLIR method, where the lexical-based method consists of two steps of translation and retrieval, and the neural-based method introduces the $\text{DISTIL}_{\text{DistilmBERT}}$ model, an end-to-end neural network. In our preliminary results, the lexical-based CLIR method performs better than the neural-based method.

Keywords: Cross-language information retrieval, Lexical-based CLIR method, Neural-based CLIR method

1 Introduction

The demand for multilingual information on the Internet is growing rapidly, and CLIR has attracted extensive attention from researchers. CLIR is the task of retrieving documents in the target language L_t with queries written in the source language L_s, which can help users retrieve information from different languages[1, 2]. The neural network models have developed rapidly and are widely used in information retrieval

tasks with advanced performance. However, CLIR does not perform very well in neural-based retrieval methods[3, 4]. The TREC 2022 NeuCLIR track proposed a neural CLIR challenge to explore the effectiveness of neural-based CLIR. In this track, the topics are written in English and the document collections are written in Chinese, Persian, and Russian respectively. For each topic, the retrieval system needs to return a ranked list of 1000 documents drawn from three document collections separately.

This paper describes our first participation in the TREC 2022 NeuCLIR track. The track comprises three tasks in total, we only focus on the Ad Hoc CLIR. In this work, we explored two approaches for CLIR. The first one is the lexical-based CLIR method. This method consists of two steps, the first step is the translation of English topics into the target language, and the second step is retrieval based on probabilistic models (TF_IDF[5], BM25[6], and PL2[7]). The second one is the neural-based CLIR method. Nowadays, multilingual text encoders pre-trained on more than 100 languages, have become the standard for multilingual representation learning and cross-language transfer in natural language processing, such as mBERT[8] or XLM[9]. However, Litschko et al. argued that pre-trained multilingual encoders produce poor sentence embeddings, resulting in lower performance on unsupervised text similarity tasks[10]. And they think encoders specialized for semantic similarity like $\text{DISTIL}_{\text{DistilmBERT}}$ are supposed to encode sentence meaning more accurately. Therefore, we use the pre-trained multilingual encoder $\text{DISTIL}_{\text{DistilmBERT}}$ to complete the end-to-end neural-based CLIR. Due to the time constraint and the limited resources, we only managed to submit three baseline results based on the PL2 model.

The remainder of this paper is structured as follows. In Section 2 we give an overview of the NeuCLIR track. Next, in Section 3 we present the implementation details of the CLIR system used in this work. And Section 4 analyzes the experimental results. Lastly, we conclude our findings and works in Section 5.

2 Tasks

The TREC 2022 NeuCLIR track includes three tasks. The main task is Ad Hoc CLIR. For each English topic, the retrieval system needs to return a ranked list of 1000 documents drawn from the target language document collections. The remaining two

tasks are: (1) a reranking task based on the Ad Hoc CLIR; (2) a monolingual retrieval task, which is the same as the Ad Hoc CLIR, except that the English topic is manually translated into the target language. An overview of the datasets provided by NeuCLIR is shown in Table 1, and all data are stored in JSONL format.

Table 1 An Overview of the Datasets Provided by NeuCLIR

Dataset Name	**Topics**		**Documents**	
	Language	Quantity	Language	Quantity
HC4-train	English	27	Chinese	646305
			Persian	486486
			Russian	4721064
NeuCLIR1	English	136	Chinese	3,179,209
			Persian	2,232,016
			Russian	4,627,543
NeuCLIR-translations	Chinese	136	En-Zho	3,179,209
	Persian		En-Fas	2,232,016
	Russian		En-Rus	4,627,543

3 Methods

3.1 The Lexical-based CLIR Method

Traditional CLIR methods use existing machine translation systems to translate documents or queries so that queries and documents are in the same language[11]. During the whole experiment, the lexical-based CLIR method we used can be summarized into two stages: translation and retrieval. Fig. 1 shows an overview of our lexical-based CLIR system. We preprocessed the data as follows. Firstly, We translated the topics into the target language by calling the Google Translate API[1]. And each topic is composed of a short title and a sentence-length description from the given dataset. Then, for the Chinese document collection, we converted traditional Chinese to

[1] http://translate.google.cn/m?q=%s&tl=%s&sl=%s

simplified Chinese with the tools provided by NeuCLIR[2]. Finally, we indexed all target language document collections by the platform PyTerrier[3]. For the retrieval step, three probabilistic models of TF_IDF, BM25, and PL2 in PyTerrier were used to conduct experiments.

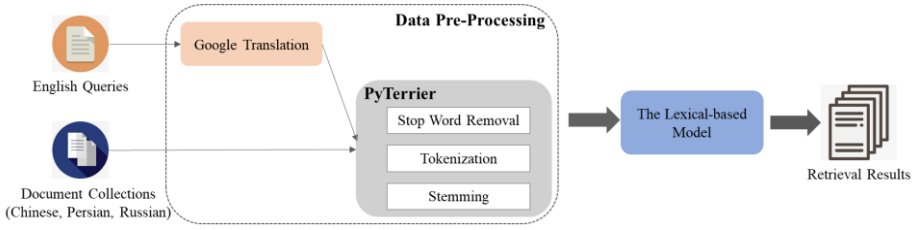

Fig. 1 The Pipeline of the Lexical-based CLIR Method

We performed a test experiment on the partial training set of the dataset HC4[12]. The preliminary test results are shown in Table 2. From the test results, the PL2 model outperformed the other two models. For this, the three baseline results we finally submitted were all obtained by the PL2 model. After the submission, we conducted experiments on the complete HC4 training data set and the results are shown in Table 3. However, the retrieval based on the PL2 model does not get the best results in this work. In addition, these results also show that the retrieval performance improves when traditional Chinese documents are converted to simplified Chinese ones.

Table 2 Test Results Retrieved by Probabilistic Models

Language	**Method**	**Metrics**			
		RR(rel=2)	AP(rel=2)	nDCG@10	nDCG@100
Chinese	TF_IDF	0.0603	0.0502	0.2231	0.3958
	BM25	0.0554	0.0483	0.2255	0.3988
	PL2	0.0790	0.0577	0.2484	0.4095

[2] https://github.com/NeuCLIR/download-collection

[3] https://pyterrier.readthedocs.io/en/latest/#

Table 3 Results Retrieved by Probabilistic Models on the Training Set of HC4 Dataset

Language	Method	Metrics			
		RR(rel=2)	AP(rel=2)	nDCG@10	nDCG@100
Chinese	TF_IDF	0.1083	0.0532	0.2447	0.3853
	BM25	0.1058	0.0506	0.2350	0.3804
	PL2	0.1068	0.0521	0.2272	0.3707
Simplified Chinese	TF_IDF	0.1082	0.0538	0.2323	0.3835
	BM25	0.1056	0.0510	0.2360	0.3816
	PL2	0.1068	0.0518	0.2278	0.3714
Persian	TF_IDF	0.3381	0.2836	0.4542	0.6102
	BM25	0.3239	0.2802	0.5017	0.6336
	PL2	0.2596	0.2580	0.4065	0.5667
Russian	TF_IDF	0.2622	0.1904	0.3340	0.4615
	BM25	0.2381	0.1702	0.2961	0.4442
	PL2	0.2286	0.1793	0.2928	0.4235

3.2 The Neural-based CLIR Method

In the neural-based method, our experiment mainly includes three steps: data preprocessing, model construction, and retrieval. Firstly, we preprocess the data which is in English, Persian, Chinese, and Russian. Then, the text is encoded using a multilingual neural model. Finally, the relevant documents are retrieved and returned. Fig. 2 shows an overview of our neural-based CLIR system. Next, the details of these three steps are described as follows.

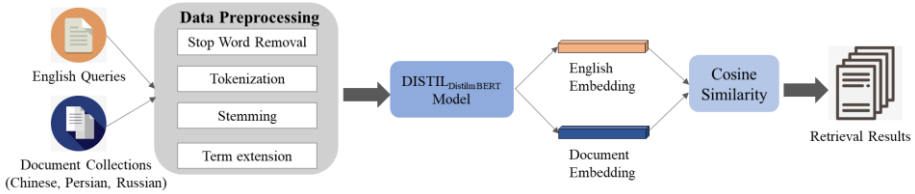

Fig. 2 The Pipeline of the Neural-based CLIR Method

Data preprocessing includes four parts: tokenization, stop word removal, stemming, and term extension. For each topic, we directly used the data provided by the task organizers and removed the invalid characters. For the document, we took the NeuCLIR1 document collections provided by the task organizers and extracted 'title' and 'detail' from the document collections as document data. For the Persian and Russian document collections, we used the NLTK[4] package. And for the Chinese document collection, we used the jieba[5] package.

The neural model $\text{DISTIL}_{\text{DistilmBERT}}$ we chose is introduced by Litschko[10]. It is a multilingual model based on knowledge distillation between the teacher-student framework. The teacher model is Sentence-BERT[12], and the student model is DsitilmBERT[13]. Sentence-BERT is a modification of pre-trained BERT, which can derive semantically meaningful sentence embeddings. The teacher model first learns professional semantic similarity knowledge, then injects this knowledge into the student model, so that the student model can simulate the output of the teacher model.

We used the trained student model to encode the topic and document respectively. After vector length normalization, the model performs retrieval by computing cosine similarity. Since the document collections are all very large and one retrieval is very time-consuming, we divide the documents into 10 copies and select the top 2000 documents for each retrieval. The final results are generated from these 20,000 documents.

4 Results and Discussion

In this work, both the lexical-based method and the neural method have experimented on the NeuCLIR1 dataset. According to the latest relevance judgments released by the TREC 2022 NeuCLIR track, the evaluations of our results using ir-measures[6] are shown in Table 4. Among them, bold represents the optimal result, and * represents the result we submitted to the organizing committee.

---

[4] http://www.nltk.org

[5] https://pypi.org/project/jieba/

[6] https://ir-measur.es/en/latest/

The results in Table 4 show that the neural-based method based on the $DISTIL_{DistilmBERT}$ model performs best in the target language Russian and worst in Persian. The reason may be that the Russian document collection has the largest number of documents and the Persian the least. And the evaluations of the results retrieved by the $DISTIL_{DistilmBERT}$ model are all lower than that of probabilistic models. And as stated in Section 3.1, the retrieval performance improves when the traditional Chinese documents are converted to simplified ones. To sum up, in this work, the lexical-based retrieval method with probabilistic models outperforms the neural-based retrieval method with $DISTIL_{DistilmBERT}$ model, and TF_IDF performed best in the probability models. Due to the time constraint and the limited resources, we only managed to submit three baseline results based on the PL2 model.

Table 4 Final Scores for Both Retrieval Methods

Language	**Method**	**Metrics**			
		map	nDCG@20	recall@100	recall@1000
Chinese	TF_IDF	**0.1060**	**0.1883**	**0.2012**	**0.3703**
	BM25	0.0920	0.1747	0.1993	0.3548
	PL2	0.0839*	0.1567*	0.1649*	0.3066*
Simplified Chinese	TF_IDF	**0.2392**	**0.3468**	**0.4848**	**0.7111**
	BM25	0.2240	0.3254	0.4686	0.6975
	PL2	0.1852*	0.2788*	0.3681*	0.6258*
	$DISTIL_{DistilmBERT}$	0.1079	0.1869	0.2348	0.4802
Persian	TF_IDF	**0.2080**	**0.3155**	**0.4850**	**0.7273**
	BM25	0.1081	0.1798	0.3583	0.5736
	PL2	0.1109*	0.1810*	0.3303*	0.5766*
	$DISTIL_{DistilmBERT}$	0.0821	0.1605	0.2946	0.5376
Russian	TF_IDF	**0.2487**	**0.3542**	**0.4482**	**0.7122**
	BM25	0.2242	0.3132	0.4207	0.6850
	PL2	0.2135*	0.2982*	0.3794*	0.6641*
	$DISTIL_{DistilmBERT}$	0.1491	0.2537	0.3066	0.5615

5 Conclusion

This paper briefly introduces what is a CLIR task and the partial information about the TREC 2022 NeuCLIR track, and then describes the work we tried with the lexical-based CLIR method and the neural-based CLIR method. Finally, according to the latest relevance judgments released by NeuCLIR, the final results of the two methods were evaluated using ir-measures. The evaluations show that the lexical-based CLIR method based on the probability models performs better than the neural-based CLIR method based on the $\text{DISTIL}_{\text{DistilmBERT}}$ model in this work.

References

1. Zhou D., Qu W., Li L., et al.: Neural Topic-enhanced Cross-lingual Word Embeddings for CLIR. Information Sciences 608: 809-824 (2022)
2. Zhou D., Truran M., Brailsford T., et al.: Translation Techniques in Cross-language Information Retrieval. ACM Computing Surveys 45: 1-44 (2012)
3. Bonab H., Sarwar S. M., Allan J.: Training Effective Neural CLIR by Bridging the Translation Gap. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Association for Computing Machinery, Virtual Event, China, 9-18 (2020)
4. Khattab O., Zaharia M.: ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Association for Computing Machinery, Virtual Event, China, 39-48 (2020)
5. Joachims T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML). Nashville, Tennessee, USA, 143-151 (1997)
6. Anick P. G., Flynn R. A.: Versioning a Full-text Information Retrieval System. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Association for Computing Machinery, Copenhagen, Denmark, 98-111 (1992)

7. Amati G., Rijsbergen C. J. V.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems 20: 357-389 (2002)
8. Devlin J., Chang M.-W., Lee K., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). Association for Computational Linguistics, Minneapolis, Minnesota, 4171-4186 (2019)
9. Conneau A., Lample G.: Cross-lingual Language Model Pretraining. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS). Vancouver, BC, Canada, 1-11 (2019)
10. Litschko R., Vulić I., Ponzetto S. P., et al.: Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval. In: Proceedings of the 43rd European Conference on IR Research (ECIR). Springer International Publishing, Online, 342-358 (2021)
11. Yu P., Allan J.: A Study of Neural Matching Models for Cross-lingual IR. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Association for Computing Machinery, Virtual Event, China, 1637-1640 (2020)
12. Lawrie D., Mayfield J., Oard D. W., et al.: HC4: A New Suite of Test Collections for Ad Hoc CLIR. In: Proceedings of the 44th European Conference on IR Research (ECIR). Springer International Publishing, Stavanger, Norway, 351-366 (2022)
13. Reimers N., Gurevych I.: Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 4512-4525 (2020)