KASYS at the TREC 2022 NeuCLIR Track

Kenya Abe
University of Tsukuba
Japan
s1911448@s.tsukuba.ac.jp

Kohei Shinden
University of Tsukuba
Japan
s2221648@s.tsukuba.ac.jp

Makoto P. Kato
University of Tsukuba
Japan
mpkato@acm.org

ABSTRACT

This paper describes the KASYS team's participation in the TREC 2022 NeuCLIR track. Our approach is *One-for-All*, which employs a single multilingual pre-trained language model to retrieve documents of any languages in response to an English query. The basic architecture is the same as ColBERT and its application to CLIR, ColBERT-X, but only a single model was trained with the mixture of MS MARCO and its translated version, neuMARCO, in our approach. Through the run submission, we evaluated two variants of the One-for-All approach, namely, the end-to-end and reranking approaches. As the first-stage retriever, the former uses approximated nearest neighbor search proposed in ColBERT, while the latter uses the track organizers' (top 1,000 documents in the baseline run were used as the results of the first-stage retrieval). To evaluate our runs, we used the results provided by the track organizers as a baseline (document translation). The official evaluation results showed that the reranking approaches outperforms the baseline in all the languages. On the other hand, the end-to-end approach achieved higher scores than the baseline only in Russian.

In addition to the submissions to the TREC 2022 NeuCLIR track, we also conducted experiments with the development data called HC4. The results in HC4 also showed a similar trend: the reranking approach was superior to the end-to-end approach in Persian and Russian. We also found the discrepancy that even in the same language, the performance of our approaches varies depending on the datasets.

KEYWORDS

CLIR, Dense retrieval, TREC NeuCLIR

1 INTRODUCTION

This paper describes the KASYS team's participation in the TREC 2022 NeuCLIR track. We submitted three runs, each of which contains rankings for three languages, i.e., Chinese, Persian and Russian. Our main purpose to participate in this track is to examine the performance of *One-for-All* approach, which employs a single model trained with resources of multiple languages, for retrieving documents of all languages[1]. The advantages of this approach include simplicity of the architecture to develop CLIR for multiple languages, and application to retrieval from mixed language collections. We also expected that training data in one language enhance the retrieval performance in the other languages, which can be seen in other NLP tasks [1].

As a baseline approach (named "KASYS-run"), we tried to reproduce ColBERT-X [8] in the NeuCLIR test collection. ColBERT-X is an extension of ColBERT [4] to CLIR, and employs a multilingual pre-trained language model for encoding queries and documents. A zero-shot learning setting was used in our run, in which the model was trained with only English training data (the MS MARCO v1 passage dataset). ColBERT-X underperformed a baseline based on BM25 with document translation in all the languages, probably since our reproduction was not successfully conducted due to a different model used as the encoder[2].

Our proposed approaches train a single multilingual pre-trained language model with training data in multiple languages. The architecture is exactly the same as ColBERT. We submitted two variants of this approach (named "KASYS_one_model" and "KASYS_onemodel-rerank"). The first variant took an end-to-end approach that embed documents in a collection, retrieve them based on approximated nearest neighbor search, and rerank the retrieved documents. The other variant retrieved documents from a collection by a first-stage retriever, which were provided by the track organizers, and then reranked them based on the ColBERT model. Thus, only the difference between the two variants is the first-stage retriever. The evaluation results showed that, in terms Recall@1000, nDCG@20 and MAP,

(1) The reranking approach showed higher scores than baseline scores in the three languages.
(2) The end-to-end approach did not perform well in terms of Recall@1000. In spite of this fact, there are some cases where the end-to-end approach achieved competitive scores in other metrics.
(3) The reranking approach outperformed the end-to-end approach in the three languages.

Additionally, we conducted experiments on HC4 (CLIR Common Crawl Collection) [6], which is development datasets of NeuCLIR. The results were similar to those at the track; the reranking approach demonstrated better performance than end-to-end approach in Persian and Russian. However, there are results which are inconsistent with those at the track; the end-to-end approach did not perform well in Russian.

In the remainder of this paper, Section 2 describes the details of our runs and Section 3 presents results and discussion. Finally, we conclude this paper with some future work in Section 4.

2 METHODOLOGY

We first introduce ColBERT-X proposed by Nair et al. [8], and then explain our approaches and submitted runs.

2.1 ColBERT-X

Our approaches are mostly based on ColBERT-X [8]. ColBERT-X extends ColBERT [5] to perform CLIR. ColBERT has a late-interaction structure, which computes the similarity between encoded representations of query and document tokens. Due to much computational

[1] A similar approach was proposed right before the run submission [7].

[2] XLM-RoBERTa-large was used in the original paper, while XLM-RoBERTa-base was used in our run.

cost, ColBERT employs a multi-stage retrieval architecture. At the first stage, candidate documents are retrieved by approximated nearest neighbor search. At the second stage, the candidate documents are reranked by the sum of maximum similarity between query and document tokens. One can also use any light-weight retrieval algorithms at the first stage and apply ColBERT only for reranking at the second stage.

Equation 1 is the score function to rerank the documents based on the similarity between query and document tokens. Let q represent a query and d represent a document. The query q is split into a sequence of tokens $q = (q_1, q_2, \ldots, q_{|q|})$ by a tokenizer. The document d is also split into a sequence of tokens $d = (d_1, d_2, \ldots, d_{|d|})$. The tokens are encoded into query and document embedding $\eta(q_i)$ and $\eta(d_j)$ by an encoder η (e.g., BERT or XLM-RoBERTa). The score of d in response to q is defined as:

$$S_{q,d} = \sum_{i=1}^{|q|} \max_{j=1,\ldots,|d|} \eta(q_i)^T \eta(d_j), \tag{1}$$

This function computes the similarity of every query-document-token pairs, getting the maximum similarity for every query token. The sum of the maximum similarity scores is used to rank the documents.

At training, ColBERT uses pairwise softmax cross-entropy as a loss function and MS MARCO triples [9] as training data. Triples consist of a query, a positive passage and negative passage. For CLIR dense retrieval, we have no sufficient data other than English for fine-tuning multilingual language models. Therefore, two training approaches were proposed by Nair et al. [8]: Zero-Shot and Translate-Train. The former approach (Zero-Shot) fine-tunes an encoder such as XLM-RoBERTa (XLM-R) and mBERT using English training data. At query time, a query is translated into the document language. Since the query and document are expressed in the same language by translation, this approach is considered as monolingual retrieval. The latter approach (Translate-Train) translates the English training data into the target language (the language used in the collection) and uses it to train the encoder. Since the training data contains English queries and documents of the target language, the trained model is expected to retrieve documents in the target language in response to English queries.

2.2 Our Approaches

As we mentioned earlier, we aim to examine the capability of *One-for-All* approach, which employs a single multilingual pre-trained language model trained with resources of multiple languages, for retrieving documents of any languages in response to an English query. This idea is similar to the Zero-Shot approach for ColBERT-X that uses a single model for all document languages. It is also similar to the Translate-Train approach in that translated resources are used for training. On the other hand, the One-for-All approach uses translated resources of multiple languages for training unlike Zero-Shot, and uses a single model for all document languages unlike Translate-Train.

In the One-for-All approach, we changed the composition of training data from monolingual triples to a mixture of triples of each language. The training data consists of MS MARCO v1 passage dataset [9], and neuMARCO, which was created by translating MS MARCO into three languages (Chinese, Persian, and Russian) with a machine translation model built on the top of Sockeye [3]. When sampling negative passages, we chose from a BM25 top-k ($k = 500$) ranked list. We trained the model with the batch size of 32 for 100,000 steps. Our implementation is based on the publicly available code of ColBERT-v1[3].

Table 1: KASYS's submitted runs.

	Encoder	Training data	First-stage
KASYS-run	XLM-R-base	English	ColBERT-X
KASYS_one_model	XLM-R-large	Multilingual	ColBERT-X
KASYS_onemodel-rerank	XLM-R-large	Multilingual	Baseline

2.3 Our Runs

We submitted three runs for TREC 2022 NeuCLIR track. They all have rankings for each of the three language, Chinese, Persian and Russian.

Table 1 summarizes the submitted runs from KASYS team. The main differences are (1) the multilingual pre-trained language model used in the run, (2) the language(s) of the training data (English (the original MS MARCO data [9]) or Multilingual (the original MS MARCO data and neuMARCO data in Chinese, Persian and Russian)), and (3) the first-stage retriever (ColBERT-X or the baseline system developed by the track organizers).

3 RESULTS AND DISCUSSION

We evaluated our runs in the TREC 2022 NeuCLIR track and with HC4 (development data in the NeuCLIR track).

3.1 TREC 2022 NeuCLIR Results

Table 2 shows the results of our runs. We use BM25 with document translation as a baseline. We computed Recall@1000, nDCG@20 and MAP scores for each language. From the results, we can see that KASYS_one_model achieved better performances in Russian, but not in Persian and Chinese when compared to the baseline. Although KASYS_one_model did not perform well at Recall@1000, for some cases (especially in Russian), this approach achieved higher scores than baseline scores in other metrics. When we focus on KASYS_onemodel-rerank, it outperformed the end-to-end approach (KASYS_one_model) in all the languages. From these results, we conclude that KASYS_onemodel-rerank is the best runs from KASYS.

3.2 HC4 Results

In addition to the run submission, we performed experiments with HC4. Table 3 shows the results of BM25 with a translated query, and the end-to-end approach and reranking approach. Note that the reranking approach reranked top-1000 documents retrieved by BM25 with a translated query, and, accordingly, Recall@1000 is identical for both runs. As was done in the HC4 paper [6], we used Patapsco [2], an open-source CLIR toolkit, to perform the BM25 retrieval. Unfortunately, however, we failed to reproduce the BM25

[3]https://github.com/stanford-futuredata/ColBERT

Table 2: Evaluation results of KASYS at the NeuCLIR.

	Recall@1000	nDCG@20	MAP
Chinese			
BM25 (document translation)	**0.7814**	0.3399	0.2636
KASYS_run	0.5249	0.2855	0.1659
KASYS_one_model	0.5628	0.3639	0.2217
KASYS_onemodel-rerank	0.7814	**0.3961**	**0.2864**
Persian			
BM25 (document translation)	**0.8292**	0.3546	0.2532
KASYS_run	0.5423	0.3101	0.1615
KASYS_one_model	0.5909	0.3304	0.2003
KASYS_onemodel-rerank	0.8292	**0.4152**	**0.2854**
Russian			
BM25 (document translation)	**0.7744**	0.2919	0.2162
KASYS_run	0.5108	0.2571	0.1509
KASYS_one_model	0.6014	0.3655	0.2256
KASYS_onemodel-rerank	0.7744	**0.4499**	**0.3205**

Table 3: Experiment results in HC4.

	Recall@1000	nDCG@100	MAP
Chinese			
BM25	0.4207	0.2051	0.1337
KASYS_one_model	**0.7359**	**0.4171**	**0.2576**
KASYS_onemodel-rerank	0.4207	0.3095	0.1996
Persian			
BM25	0.7587	0.3535	0.2272
KASYS_one_model	**0.8207**	0.4443	0.2729
KASYS_onemodel-rerank	0.7587	**0.4531**	**0.2897**
Russian			
BM25	**0.7104**	0.3469	0.2233
KASYS_one_model	0.6008	0.2856	0.1792
KASYS_onemodel-rerank	**0.7104**	**0.3694**	**0.2282**

results of the original paper especially in Chinese. Possibly due to this problem, the reranking approach (KASYS_onemodel_rerank) did not perform in Chinese, althogh the reranking approach makes better results than baseline results (First-stage). Whereas, in Persian and Russian, the reranking approach outperformed the end-to-end approach (KASYS_one_model). This finding is consistent with that in the NeuCLIR track (see Section 3.1). The end-to-end approach (KASYS_one_model) showed better performances than the baseline in Chinese and Persian, but underperformed in Russian. This finding is inconsistent with the trend observed in Table 2.

4 CONCLUSIONS

In this paper, we introduced the One-for-All approach for CLIR, which uses only a single multilingual pre-trained language model trained with resources of multiple languages, for retrieving documents of any languages in response to an English query. Two variants of the One-for-All approach were evaluated in the NeuCLIR track and HC4, namely, the end-to-end and reranking approaches. The experimental results, though preliminary, showed the following findings:

(1) The reranking approaches achieved much higher performances than the NeuCLIR baseline in the three languages. In HC4, the reranking approaches also worked well in Persian and Russian.
(2) For NeuCLIR, the end-to-end approach performed well in Russian but not in Chinese and Russian. In HC4, an opposite trend was found: it was better than the baseline in Chinese and Persian but not in Russian.
(3) The reranking approach outperformed the end-to-end approach in many cases: all of the three languages in NeuCLIR, and Persian and Russian in HC4.

We observed that our approach has a weakness at addressing some languages depending on the datasets. To study in more detail, we will compare our approaches with the model trained only with English. Another important finding is that the reranking approach outperformed the end-to-end approach. We hypothesize that the results might be different in different training settings (e.g., the negative sampling method and the number of triples). We will analyze the above two issues in our future work.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP22H03905 and JP21H03554.

REFERENCES

[1] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.
[2] Cash Costello, Eugene Yang, Dawn Lawrie, and James Mayfield. 2022. Patapasco: a Python framework for cross-language information retrieval experiments. In *European Conference on Information Retrieval*. Springer, 276–280.
[3] Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The sockeye 2 neural machine translation toolkit at AMTA 2020. *arXiv preprint arXiv:2008.04885* (2020).
[4] Omar Khattab and Matei Zaharia. 2020. 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*. 39–48.
[5] Omar Khattab and Matei Zaharia. 2020. 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*. 39–48.
[6] Dawn Lawrie, James Mayfield, Douglas W Oard, and Eugene Yang. 2022. HC4: a new suite of test collections for ad hoc CLIR. In *European Conference on Information Retrieval*. Springer, 351–366.
[7] Dawn Lawrie, Eugene Yang, Douglas W Oard, and James Mayfield. 2022. Multilingual ColBERT-X. *arXiv preprint arXiv:2209.01335* (2022).
[8] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*. Springer, 382–396.
[9] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*.