Learning Cross-Lingual IR from an English Retriever

Yulong Li[†*], **Martin Franz**[‡*], **Md Arafat Sultan**[‡*],
Bhavani Iyer[‡], **Young-Suk Lee**[‡] and **Avirup Sil**[‡]
[†]IBM Research, [‡]IBM Research AI
{yulongl, franzm, bsiyer, ysuklee, avi}@us.ibm.com
arafat.sultan@ibm.com

Abstract

We present a new cross-lingual information retrieval (CLIR) system trained using multi-stage knowledge distillation (KD). The teacher relies on a highly effective but expensive two-stage process consisting of query translation and monolingual IR, while the student executes a single CLIR step. We teach the student powerful multilingual encoding as well as CLIR by optimizing two corresponding KD objectives. Learning useful non-English representations from an English-only retriever is accomplished through a cross-lingual token alignment algorithm that relies on the representation capabilities of the underlying multilingual language model. In both in-domain and zero-shot evaluation, the proposed method demonstrates far superior accuracy over direct fine-tuning with labeled CLIR data. One of our systems is also the current best single-model system on the XOR-TyDi leaderboard.

1 Introduction

Multilingual models are critical for the democratization of AI. Cross-lingual information retrieval (CLIR) (Braschler et al., 1999; Shakery and Zhai, 2013; Jiang et al., 2020; Asai et al., 2021a), for example, can find relevant text in a high-resource language such as English even when the query is posed in a different, possibly low-resource, language. In this work, we develop useful CLIR models for this constrained, yet important, setting where a retrieval corpus is available only in a single high-resource language (English in our experiments).

A straightforward solution to this problem can be based on machine translation (MT) of the query into English, followed by English IR (Asai et al., 2021a). While this two-stage process is capable of providing accurate predictions, an alternative end-to-end approach that can tackle the problem purely cross-lingually, *i.e.*, without involving MT, would clearly be more efficient and cost-effective. Pre-trained multilingual masked language models (PLMs) such as multilingual BERT (Devlin et al., 2019) or XLM-RoBERTa (XLM-R) (Conneau et al., 2020) can provide the foundation for such an approach, as one can simply fine-tune a PLM with labeled CLIR data (Asai et al., 2021b).

We first run an empirical evaluation of these two approaches on a public CLIR benchmark (Asai et al., 2021a), which includes both in-domain and zero-shot out-of-domain tests. We use ColBERT (Khattab and Zaharia, 2020; Khattab et al., 2021) as our IR architecture[1] and XLM-R as the underlying PLM for both methods (§2). Results indicate that the MT-based solution can be vastly more effective than CLIR fine-tuning, with observed differences in Recall@5kt of 22.2–28.6 points (§3). Crucially, the modular design of the former allows it to leverage additional English-only training data for its IR component, providing significant boosts to its results.

The above findings lead naturally to the central research question of this paper: Can a high-performance CLIR model be trained that can operate without having to rely on MT? To answer the question, instead of viewing the MT-based approach as a competing one, we propose to leverage its strength via knowledge distillation (KD) into an end-to-end CLIR model. KD (Hinton et al., 2014) is a powerful supervision technique typically used to distill the knowledge of a large *teacher* model about some task into a smaller *student* model (Mukherjee and Awadallah, 2020; Turc et al., 2020). Here we propose to use it in a slightly different context, where the teacher and the student retriever are identical in size, but the former has superior performance simply due to utilizing MT output and consequently operating in a high-resource and low-difficulty monolingual environment.

[*] Equal contribution.

[1] Due to its state-of-the-art (SOTA) performance, outperforming recent models such as DPR (Karpukhin et al., 2020).

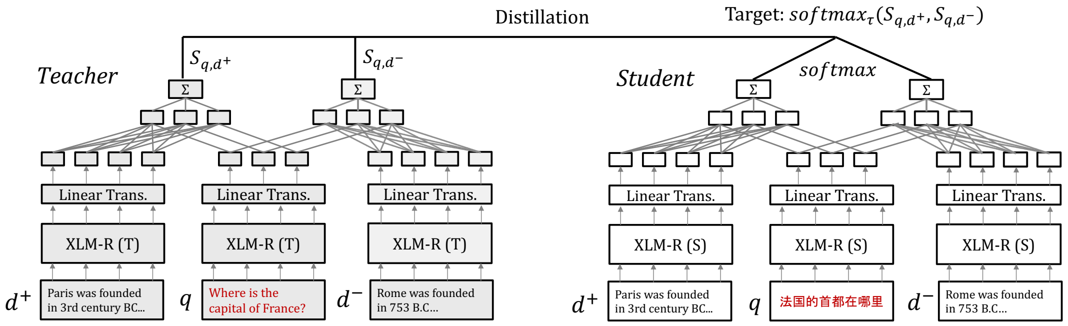

Figure 1: Relevance score distillation. The teacher is shown all-English triples while the student's query input is non-English. Training minimizes the student's KL divergence from the teacher's output softmax distribution over S_{q,d^+} and S_{q,d^-} (τ is the temperature).

We run two independent KD operations (§2.2). One directly optimizes an IR objective by utilizing labeled CLIR data: parallel questions (English and non-English) and corresponding relevant and non-relevant English passages. The teacher and the student are shown the English and non-English versions of the questions, respectively; the training objective is for the student to match the soft query-passage relevance predictions of the teacher. The second KD task is representation learning from parallel text, where the student learns to encode a non-English text in a way that matches the teacher's encoding of the aligned English text, *at the token level*. The cross-lingual token alignment needed to create the training data for this task is generated using a greedy alignment process that exploits the PLM's multilingual representations.

In our experiments on the XOR-TyDi dataset (Asai et al., 2021a), the KD student outperforms the fine-tuned ColBERT baseline by 25.4 (in-domain) and 14.9 (zero-shot) Recall@5kt, recovering much of the performance loss from the MT-based solution. It is also the best single-model system on the XOR-TyDi leaderboard[2] at the time of writing. Ablation studies show that each of our two KD processes contribute significantly towards the final performance of the student model.

Our contributions can be summarized as follows: **(1)** We present an empirical study of the effectiveness of a SOTA IR method (ColBERT) on cross-lingual IR with and without MT. **(2)** We propose a novel end-to-end cross-lingual solution that uses knowledge distillation to learn both improved text representation and retrieval. **(3)** We demonstrate with a new cross-lingual alignment algorithm that distillation using parallel text can strongly augment cross-lingual IR training. **(4)** We achieve new single-model SOTA results on XOR-TyDi.

2 Method

Here we first describe our base IR architecture (ColBERT) and then the proposed KD-based cross-lingual training algorithms.

2.1 The ColBERT Model

ColBERT (Khattab and Zaharia, 2020) employs a transformer-based encoder to separately encode the input query and document, followed by a linear compression layer. Each training instance is a <q, d^+, d^-> triple, where q is a query, d^+ is a positive (relevant) document and d^- is a negative (non-relevant) document. A relevance score $S_{q,d}$ for the pair (q, d) is computed using Eq. 1, where $d \in \{d^+, d^-\}$ and E_{q_i} and E_{d_j} are the output embeddings of query token q_i and document token d_j, respectively. For a given training triple, a cross-entropy loss is minimized for the S_{q,d^+} and S_{q,d^-}.

$$S_{q,d} := \sum_{i \in [|q|]} max_{j \in [|d|]} E_{q_i} \cdot E_{d_j}^T \qquad (1)$$

For inference, the embeddings of all documents are calculated *a priori*, while the query embeddings and the relevance score are computed at runtime.

2.2 Knowledge Distillation

Our teacher and student are both ColBERT models that fine-tune the same underlying multilingual PLM for IR. The teacher is first trained with all-English triples using the procedure of §2.1. The goal of the subsequent KD training is to teach the student how to reproduce the behavior of this teacher, provided that it sees non-English questions while the teacher sees English questions.

[2] https://nlp.cs.washington.edu/xorqa/

```
Input:
$v_T$: Teacher's output representation of tokenized
  English (EN) text.
$v_S$: Student's representation of parallel non-EN text.
Output:
$v_T^{(a)}$: Reordered teacher output embeddings to reflect
  position-wise alignment with $v_S$.
Procedure:
$DM \leftarrow cosine\_distance(v_T, v_S)$ //matrix
//get index pairs to swap in $v_T$
$swaps \leftarrow [\,]$
for row in rows(DM) do
    //loop runs $|v_T|$ times
    $minValue \leftarrow min(DM)$
    $i, j \leftarrow index\_of(minValue)$
    //swap rows i and j
    $DM[[i, j], :] = DM[[j, i], :]$
    //set row $j$ and column $j$ to $+\infty$
    $DM[j, :] \leftarrow +\infty$
    $DM[:, j] \leftarrow +\infty$
    $swaps.append((i, j))$
end
//swap teacher's output tokens
$v_T^{(a)} \leftarrow v_T$
for s in swaps do
    $v_T^{(a)}[s[0], s[1]] \leftarrow v_T[s[1], s[0]]$
end
```

Algorithm 1: Cross-lingual alignment.

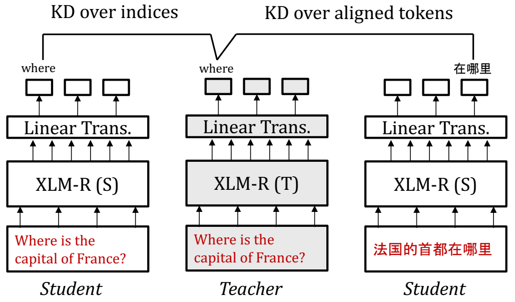

Figure 2: Distillation for representation learning. The student learns to encode both English and non-English tokens in context that matches the teacher's output embeddings for corresponding English tokens.

We apply KD at two different stages of the ColBERT workflow: (1) relevance score computation ($S_{q,d}$ in Eq. 1), and (2) encoding (*e.g.*, E_{q_i}). Figure 1 depicts the former in detail, where training minimizes the KL divergence between the student's and the teacher's output softmax distributions (with temperature) over S_{q,d^+} and S_{q,d^-}.

Labeled training data for CLIR are of limited availability, whereas MT, being a more established area of research, has produced a large amount of parallel text over the years. We seek to exploit parallel corpora in our second KD training stage, where we train the student to compute representations for non-English texts that closely match the teacher's representations of aligned English texts. Crucially, since ColBERT computes a single vector for each individual input token (*i.e.*, a PLM vocabulary item) and not for the entire input text, our algorithm must support distillation at the token level.

To achieve this, we apply an unsupervised cross-lingual alignment algorithm. Assuming $(ne_1, ..., ne_S)$ to be the ordered tuple of tokens in a non-English text and $(e_1, ..., e_T)$ the corresponding tuple from the aligned English text, each iteration of this algorithm greedily aligns the next (ne_i, e_j) pair with the highest cosine similarity of their output embeddings. Algorithm 1 implements this idea by repositioning the teacher's tokens so that they are position-wise aligned with the corresponding student tokens. Note that the design choice of using a common multilingual PLM in the teacher and the student, even though the former is tasked only with handling English content, is key for the operation of this algorithm as it relies on the PLM's multilingual representation capabilities.

In addition to cross-lingual alignment, we also perform a similar KD procedure in which both the teacher and the student are shown the same English text. This step is useful because ColBERT uses a shared encoder for the query and the document, necessitating a student that is able to effectively encode text from English documents as well.

Using the alignment information, we train the student by minimizing the Euclidean distance between its representation of a token (English or non-English) and the teacher's representation of the corresponding English token. Figure 2 shows the KD process for representation learning.

3 Experiments

Setup. Our primary CLIR dataset is XOR-TyDi (Asai et al., 2021a), which contains examples in seven typologically diverse languages: Arabic (Ar), Bengali (Bn), Finnish (Fi), Japanese (Ja), Korean (Ko), Russian (Ru) and Telugu (Te). For standard in-domain experiments, we use a train-dev-test split of this dataset. There are 2,113 questions in the test set. For zero-shot experiments, we use the MKQA (Longpre et al., 2020) dataset for training and validation, and the following shared languages in the XOR-TyDi test set for evaluation: Ar, Fi, Ja, Ko and Ru. Both training sets contain English questions and their human translations in the other languages, their short answers and corresponding relevant (positive) and non-relevant (neg-

System	R@5kt
With target-domain supervision:	
$\text{ColBERT}_{\text{CL}}$: *ft*(XOR)	32.9
$\text{ColBERT}_{\text{EN+CL}}$: *ft*(NQ) → *ft*(XOR)	47.7
Teacher: MT + $\text{ColBERT}_{\text{EN}}$	76.3
$\text{ColBERT}_{\text{EN+CL}} \rightarrow \text{KD}_{\text{PC}} \rightarrow \text{KD}_{\text{XOR}}$	73.1
Zero-shot:	
$\text{ColBERT}_{\text{CL}}$: *ft*(MKQA)	23.6
$\text{ColBERT}_{\text{EN+CL}}$: *ft*(NQ) → *ft*(MKQA)	46.9
Teacher: MT + $\text{ColBERT}_{\text{EN}}$	69.1
$\text{ColBERT}_{\text{EN+CL}} \rightarrow \text{KD}_{\text{PC}} \rightarrow \text{KD}_{\text{MKQA}}$	61.8

Table 1: Performance on XOR-TyDi. *ft*: fine-tuning; NQ: the OpenNQ dataset; PC: parallel corpus. Direct fine-tuning of ColBERT with IR triples underperforms MT + English IR by 22.2–28.6 points; our KD-based method closes this gap by 67.1%–88.8%.

System	R@5kt
With target-domain supervision:	
$\text{ColBERT}_{\text{EN+CL}} \rightarrow \text{KD}_{\text{PC}} \rightarrow \text{KD}_{\text{XOR}}$	73.1
$\text{ColBERT}_{\text{EN+CL}} \rightarrow \text{KD}_{\text{PC}}$	68.6
$\text{ColBERT}_{\text{EN+CL}} \rightarrow \text{KD}_{\text{XOR}}$	63.6
$\text{ColBERT}_{\text{EN+CL}}$	47.7
Zero-shot:	
$\text{ColBERT}_{\text{EN+CL}} \rightarrow \text{KD}_{\text{PC}} \rightarrow \text{KD}_{\text{MKQA}}$	61.8
$\text{ColBERT}_{\text{EN+CL}} \rightarrow \text{KD}_{\text{PC}}$	55.9
$\text{ColBERT}_{\text{EN+CL}} \rightarrow \text{KD}_{\text{MKQA}}$	49.3
$\text{ColBERT}_{\text{EN+CL}}$	46.9

Table 2: Ablation study results. KD with parallel corpus (KD_{PC}) and IR triples (KD_{XOR}) both play key roles in our student model. Interestingly, the former has a more positive impact on its performance.

ative) Wikipedia snippets. Additionally, we use the OpenNQ (Kwiatkowski et al., 2019) training set for English pre-training of the baseline model. Further details on data pre-processing, the final training sets and the optimal hyperparameter configurations are provided in Appendix A.1 and A.2.

The CLIR baseline used in our experiments is a ColBERT model with an XLM-R PLM, which we iteratively fine-tune first on English and then on cross-lingual IR triples for optimal performance. Our student model is initialized with the parameter weights of this baseline, and is further fine-tuned using the two KD objectives. As the KD teacher, we train a model on only English triples, as stated before. During evaluation, it is given machine-translated questions that come with the XOR-TyDi dataset. Appendix A.1 contains more details on the supervision of these models.

Our evaluation metric is Recall at 5000 tokens (R@5kt) (Asai et al., 2021a), which computes the fraction of questions for which the ground truth short answer is contained within the top 5,000 tokens of the retrieved passages. Appendix A.3 also presents R@2kt results.

Evaluation. Table 1 compares the performance of our models measured by their R@5kt scores. First, we observe that pre-training the baseline model on English IR triples from the OpenNQ dataset (Kwiatkowski et al., 2019) substantially boosts its performance in both in-domain and zero-shot settings. However, it still underperforms the MT + English IR pipeline by 28.6 and 26.5 points, respectively. The proposed KD-based procedure, first with the parallel corpus (for representation learning) and then with the IR triples (for CLIR), achieves an overall improvement of 25.4 points for the baseline model in in-domain evaluation, which, quite impressively, is only 3.2 points behind the teacher's score. In the zero-shot setting, we again observe a sizable gain of 14.9 points. We present more fine-grained per-language results in Appendix A.4.

Leaderboard Submission. Our KD student of Table 4 row 4, trained on XOR-TyDi and submitted under the name CHIMAERAS B, is the best single-model system on the official XOR-TyDi leaderboard at the time of this writing. Please see Appendix A.5 for further details.

Ablation Study. We train two more student models, each of which goes through only one of the two KD training steps. Table 2 summarizes the results: KD with only CLIR examples and with only the parallel corpus improves the system's score by 15.9 and 20.9 points, respectively with target-domain supervision. Interestingly, although the parallel corpus does not provide any IR signal, it contributes more to the final performance of the model. These results also confirm that our cross-lingual alignment algorithm indeed produces useful alignments.

4 Conclusion

We train highly effective end-to-end cross-lingual IR models by distilling the knowledge of an English retriever. We propose separate processes to teach IR and multilingual text representations, and present for the latter a cross-lingual alignment algorithm that only relies on the underlying masked language model's multilingual representation capabilities. Supervised and zero-shot evaluations show that our model recovers much of the performance lost due to operating in an efficient cross-lingual mode. Our KD-based method also yields new single-model SOTA results on the XOR-TyDi benchmark.

5 Ethics

5.1 Limitations

We show the effectiveness of multi-stage knowledge distillation and cross-lingual token alignment in training a cross-lingual information retrieval system. We believe that it can be transferred to more datasets and languages, but here we only show proof of concept for the XOR-TyDi and MKQA datasets and the seven languages mentioned in the manuscript.

5.2 Risks

The intent of this work is to develop a new method for high-performance cross-lingual information retrieval. It is possible that a malicious user could try to attack the system by providing poor or offensive training data. We do not support it being used in such a manner. The risks of our system are the same as other NLP systems and we do not believe we introduce any additional risk.

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual Open-Retrieval Question Answering. In *NAACL*.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. In *NeurIPS*.

Martin Braschler, Jürgen Krause, Carol Peters, and Peter Schäuble. 1999. Cross-language information retrieval (clir) track overview. In *TREC*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning Workshop*.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with bert. *arXiv preprint arXiv:2004.13005*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for openqa with colbert. *Transactions of the ACL*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *SIGIR*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *arXiv preprint arXiv:2007.15207*.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. XtremeDistil: Multi-stage Distillation for Massive Multilingual Models. In *ACL*.

Azadeh Shakery and ChengXiang Zhai. 2013. Leveraging comparable corpora for cross-lingual information retrieval in resource-lean language pairs. *Information retrieval*, 16(1):1–29.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2020. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. In *ICLR*.

A Appendix

A.1 Data Pre-processing

The official XOR-TyDi training set consists of 15,221 natural language queries, their short answers, and examples of corresponding relevant (positive) and non-relevant (negative) Wikipedia snippets. For most queries, there are one positive and three negative examples. We remove the 1,699 (11%) questions that have no answers in the dataset. A random selection of 90% of the remaining examples is used for training and the rest for validation.

Following the original XOR-TyDi process, we also obtain additional training examples by running BM25-based retrieval against a Wikipedia corpus and using answer string match as the relevance criterion. These examples are added to the original set to obtain three positive and 100 negative examples per query. As the blind test set for final evaluation, we use the 2,113 questions in the official XOR-TyDi dev set.

Our monolingual (English) training data of about 17.5M triples are derived from the third fine-tuning round (ColBERT-QA3) of ColBERT relevance-guided supervision (Khattab et al., 2021) with OpenNQ data (Kwiatkowski et al., 2019). The parallel corpus used in our KD experiments for representation learning is constructed from three different sources: (1) an in-house crawl of Korean, (2) LDC releases (Arabic), and (3) OPUS[3]. The corpus has a total of 6.9M passage pairs which include .9M pairs in Telugu and 1M pairs in each of the other six languages.

For zero-shot experiments, the training examples are derived from MKQA (Longpre et al., 2020), which consists of 10k queries selected from NQ, human translated into 25 additional languages, five of which overlap with XOR-TyDI: Ar, Fi, Ja, Ko and Ru. We construct training data (triples) from 2,037 queries translated into these five languages for which there are corresponding positive and negative passages in the OpenNQ dataset. For each of the five languages, there are 519k triples for a total of 2.6M triples. We set aside 200 queries translated into the 5 languages for a total of 1,000 queries as a development set. We remove all MKQA queries from the OpenNQ training data for these experiments.

The CLIR baseline for our experiments is a ColBERT model with an XLM-R PLM, which we first fine-tune with 17.5M NQ examples for one epoch and then 2.9M XOR-TyDi triples for five epochs. Our student model is initialized with the parameter weights of the baseline, and is further fine-tuned using the two KD objectives. The monolingual teacher model—also a ColBERT model running on top of the pre-trained XLM-R—is trained with only the 17.5M NQ triples for one epoch.

A.2 Model Selection

All the models were trained with single Nvidia A100 GPU. The longest training time for a single model was less than 200 hours. Following are the final hyperparameter configurations of our different models. They were selected based on the respective validation sets performance.

Hyperparameter	Value
Standard ColBERT hyperparameters:	
batch size	192
gradient accumulation steps	6
linear compression dim	128
query maxlen	32
document maxlen	180
Target-domain supervision	
Baseline model:	
lr (NQ)	1.5e-6
lr (XOR)	6e-6
# Epochs (NQ)	1
# Epochs (XOR)	5
Knowledge distillation:	
loss function (XOR)	KLDiv
loss function (Parallel corpus)	MSE
KD temperature (XOR)	2
lr (XOR)	6e-6
lr (Parallel corpus)	4.8e-5
# Epochs (XOR)	5
# Epochs (Parallel corpus)	2
Zero-shot	
Baseline model:	
lr (NQ)	1.5e-6
lr (MKQA)	6e-6
# Epochs (NQ)	1
# Epochs (MKQA)	1
Knowledge distillation:	
loss function (MKQA)	KLDiv
loss function (Parallel corpus)	MSE
KD temperature (MKQA)	1
lr (MKQA)	3e-7
lr (Parallel corpus)	2.4e-5
# Epochs (MKQA)	1
# Epochs (Parallel corpus)	2

Table 3: Hyperparameters used in our test set runs.

A.3 R@2kt Results

In Table 4, we present results for R@2kt which evaluates a smaller subset of the top retrievals than R@5kt. The general pattern in these results is very

[3] https://opus.nlpl.eu

System	R@2kt
With target-domain supervision:	
ColBERT$_{\text{CL}}$: *ft*(XOR)	23.9
ColBERT$_{\text{EN+CL}}$: *ft*(NQ) → *ft*(XOR)	38.1
Teacher: MT + ColBERT$_{\text{EN}}$	70.5
ColBERT$_{\text{EN+CL}}$ → KD$_{\text{PC}}$ → KD$_{\text{XOR}}$	66.0
Zero-shot:	
ColBERT$_{\text{CL}}$: *ft*(MKQA)	16.7
ColBERT$_{\text{EN+CL}}$: *ft*(NQ) → *ft*(MKQA)	38.7
Teacher: MT + ColBERT$_{\text{EN}}$	62.7
ColBERT$_{\text{EN+CL}}$ → KD$_{\text{PC}}$ → KD$_{\text{MKQA}}$	54.3

Table 4: Performance on XOR-TyDi in terms of R@2kt. *ft*: fine-tuning; NQ: the OpenNQ dataset; PC: parallel corpus; KD: knowledge distillation.

System	R@2kt
With target-domain supervision:	
ColBERT$_{\text{EN+CL}}$ → KD$_{\text{PC}}$ → KD$_{\text{XOR}}$	66.0
ColBERT$_{\text{EN+CL}}$ → KD$_{\text{PC}}$	60.6
ColBERT$_{\text{EN+CL}}$ → KD$_{\text{XOR}}$	56.6
ColBERT$_{\text{EN+CL}}$	38.1
Zero-shot:	
ColBERT$_{\text{EN+CL}}$ → KD$_{\text{PC}}$ → KD$_{\text{MKQA}}$	54.3
ColBERT$_{\text{EN+CL}}$ → KD$_{\text{PC}}$	47.7
ColBERT$_{\text{EN+CL}}$ → KD$_{\text{MKQA}}$	40.9
ColBERT$_{\text{EN+CL}}$	38.7

Table 5: Ablation study results for R@2kt.

similar to that for R@5kt: the student model closes much of the performance gap between the teacher and the baseline in both settings. Table 5 also shows the R@2kt results of the ablation study, which is again similar to the R@5kt results of Table 2.

A.4 Results on Individual Languages

Table 6 compares the performance (R@5kt) of our KD student with that of the baseline CLIR model on each individual language. In both settings, *i.e.*, with and without target-domain supervision, the student consistently outperforms the baseline, with large gains across all languages. These results point to the robustness of our approach, stemming from the individual strengths of MT, English IR and KD.

A.5 Leaderboard Submission

We anonymously submitted the official XOR-TyDi test set predictions of the KD student of Table 1 row 4 to the XOR-TyDi leaderboard[4], under the model name CHIMAERAS B. Since our parallel corpus extraction process relies on in-house source code that is not publicly available, we submitted to the "Systems using External APIs" category. It should be noted here though that our approach is not dependent on any specific extraction algorithm; we used an existing parallel corpus originally created for a different purpose simply for convenience. CHIMAERAS B, at the time of this writing, is the best single-model system on the leaderboard, outperformed only by a ColBERT ensemble that also uses MT. Crucially, all other submitted systems under the External APIs category rely on MT at decoding time, avoiding which is one of the primary goals of our work.

Language	Baseline	KD Student
Target-domain supervision:		
Te	63.0	83.2
Bn	53.3	85.9
Fi	49.4	69.4
Ja	39.4	65.1
Ko	44.9	68.8
Ru	39.2	68.8
Ar	44.3	70.2
Avg	**47.7**	**73.1**
Zero-shot:		
Fi	55.4	66.9
Ja	44.0	58.5
Ko	48.4	62.8
Ru	41.4	57.8
Ar	45.3	61.8
Avg	**46.9**	**61.8**

Table 6: R@5kt scores for in-domain and zero-shot evaluation on individual languages. Baseline for the target-domain experiment: ColBERT$_{\text{EN+CL}}$: *ft*(NQ) → *ft*(XOR). Baseline for the zero-shot experiment: ColBERT$_{\text{EN+CL}}$: *ft*(NQ) → *ft*(MKQA)

A.6 Qualitative Analysis

To find out what exact weaknesses of the baseline model the proposed method helps to address, we examine thirty random zero-shot test examples where the baseline fails to retrieve the correct answer in the top 5k tokens, but the KD student succeeds within the top 3 passages. We show four examples in Table 7 with human translations of the original non-English questions. The vast majority of our observed cases are related to weak cross-lingual encoding on the baseline model's part, where at least one important non-English word/entity in the question seems to be incorrectly matched with a similar but different English entity in the passage (e.g., the name of a different place). For the Korean, Russian and Arabic queries in the table, we observe the presence of such topically similar entities (*e.g.*, *microwave* ↔ *gamma-ray*, *Germany* ↔ places in North America). Much more rarely, we see cases similar to the Japanese query where the retrieved passage is completely off-topic.

[4] https://nlp.cs.washington.edu/xorqa/

Korean Query: 우주 마이크로파 배경은 누가 처음 발견했나요?

English Translation: Who first discovered the cosmic microwave background?

Answer: Arno Penzias and Robert Wilson

Baseline Top Passage: History of gamma-ray burst research The history of gamma-ray began with the serendipitous detection of a gamma-ray burst (GRB) on July 2, 1967, by the U.S. Vela satellites. After these satellites detected fifteen other GRBs, Ray Klebesadel of the Los Alamos National Laboratory published the first paper on the subject, "Observations of Gamma-Ray Bursts of Cosmic Origin". As more and more research was done on these mysterious events, hundreds of models were developed in an attempt to explain their origins. Gamma-ray bursts were discovered in the late 1960s by the U.S. Vela nuclear test detection satellites. The Velas were built

KD ***Student Top Passage:*** project to date, it consists of more than 66 dishes and was built in an international collaboration by Europe, North America, East Asia and Chile. A major recent focus of microwave radio astronomy has been mapping the cosmic microwave background radiation (CMBR) discovered in 1964 by radio astronomers Arno Penzias and Robert Wilson. This faint background radiation, which fills the universe and is almost the same in all directions, is "relic radiation" from the Big Bang, and is one of the few sources of information about conditions in the early universe. Due to the expansion and thus cooling of the

Russian Query: В каком сражении участвовал крейсер "Аврора"?

English Translation: What battle did the cruiser Aurora take part in?

Answer: Battle of Tsushima

Baseline Top Passage: Battle of the Arar The Battle of the Arar was fought between the migrating tribes of the Helvetii, and four Roman legions (Legions VII, VIII, IX "Hispana" and X "Equestris"), under the command of Gaius Julius Caesar, in 58 BC. This was the first major battle of the Gallic Wars. The Helvetii were a tribe that originated from what is now Switzerland. Just prior to the battle with Caesar, they had commenced on a mass migration through Roman Gaul towards the Atlantic coast. At Geneva, the Romans destroyed the wooden bridge across the Rhone and constructed 19 mi of fortifications.

KD ***Student Top Passage:*** ordered back to Asia as part of the Russian 2nd Pacific Squadron, a collection formed from the Russian Baltic Fleet, under the command of Vice-Admiral Zinovy Rozhestvensky. On the way to the Far East, "Aurora" sustained light damage from confused friendly fire, which killed the ship's chaplain and a sailor, in the Dogger Bank incident. On 27 and 28 May 1905 "Aurora" took part in the Battle of Tsushima, along with the rest of the Russian squadron. During the battle her captain, Captain 1st rank Eugene R. Yegoryev, and 14 crewmen were killed. The executive officer, Captain of 2nd rank

Japanese Query: 抑された記憶を提唱したのは誰

English Translation: Who proposed repressed memory

Answer: Sigmund Freud

Baseline Top Passage: 14 on the charts on another reissue at the end of 1979, and is now regarded as the Moody Blues signature song by British audiences. In the US, "Nights in White Satin" did not make the "Billboard" Hot 100 in 1968, although it reached No. 2 on re-release in 1972; "Tuesday Afternoon" was more successful on initial release stateside, peaking at No. 24 on the Billboard Hot 100. The 1968 follow-up LP, "In Search of the Lost Chord" included "Legend of a Mind", a song written by Ray Thomas in tribute to LSD guru Timothy Leary which encompassed a flute

KD ***Student Top Passage:*** literary example of traumatic memory that has been repressed by an otherwise healthy individual, and then recovered." Pope claimed that no entrant had satisfied the criteria. Ross Cheit, a political scientist at Brown University, cited Nina, a 1786 opera by the French composer Nicolas Dalayrac. The concept of repressed memory originated with Sigmund Freud in his 1896 essay "Zur Ätiologie der Hysterie" ("On the etiology of hysteria"). One of the studies published in his essay involved a young woman by the name of Anna O. Among her many ailments, she suffered from stiff paralysis on the right side of her

Arabic Query: ما هو اكبر اقاليم المانيا؟

English Translation: What is the largest region of Germany?

Answer: Bavaria

Baseline Top Passage: the original name of Montana was adopted. Montana is one of the nine Mountain States, located in the north of the region known as the Western United States. It borders North Dakota and South Dakota to the east. Wyoming is to the south, Idaho is to the west and southwest, and three Canadian provinces, British Columbia, Alberta, and Saskatchewan, are to the north. With an area of , Montana is slightly larger than Japan. It is the fourth largest state in the United States after Alaska, Texas, and California; it is the largest landlocked U.S. state. The state's topography is

KD ***Student Top Passage:*** Bavaria (; German and Bavarian: "Bayern" ;), officially the Free State of Bavaria (German and Bavarian: "Freistaat Bayern"), is a landlocked federal state of Germany, occupying its southeastern corner. With an area of 70,550.19 square kilometres (27,200 sq mi), Bavaria is the largest German state by land area. Its territory comprises roughly a fifth of the total land area of Germany. With 13 million inhabitants, it is Germany's second-most-populous state after North Rhine-Westphalia. Bavaria's capital and largest city, Munich, is the third-largest city in Germany. The history of Bavaria stretches from its earliest settlement and formation as

Table 7: Examples of cases where the baseline model fails to retrieve a relevant passage but the KD student succeeds within top 3. We only show the top retrieval for each system. Most errors are related to potential word/entity mistranslations, the only exception being the Japanese query where the issue is a weaker understanding of the passage content.