C3: Continued Pretraining with Contrastive Weak Supervision for Cross Language Ad-Hoc Retrieval

Eugene Yang
HLTCOE,
Johns Hopkins University, USA
eyang35@jhu.edu

Suraj Nair
University of Maryland, USA
srnair@umd.edu

Ramraj Chandradevan
Emory University, USA
rchand31@emory.edu

Rebecca Iglesias-Flores
University of Pennsylvania, USA
irebecca@seas.upenn.edu

Douglas W. Oard
University of Maryland, USA
oard@umd.edu

ABSTRACT

Pretrained language models have improved effectiveness on numerous tasks, including ad-hoc retrieval. Recent work has shown that continuing to pretrain a language model with auxiliary objectives before fine-tuning on the retrieval task can further improve retrieval effectiveness. Unlike monolingual retrieval, designing an appropriate auxiliary task for cross-language mappings is challenging. To address this challenge, we use comparable Wikipedia articles in different languages to further pretrain off-the-shelf multilingual pretrained models before fine-tuning on the retrieval task. We show that our approach yields improvements in retrieval effectiveness.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**.

KEYWORDS

cross-language information retrieval, neural methods, pretraining

ACM Reference Format:
Eugene Yang, Suraj Nair, Ramraj Chandradevan, Rebecca Iglesias-Flores, and Douglas W. Oard. 2022. C3: Continued Pretraining with Contrastive Weak Supervision for Cross Language Ad-Hoc Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531886

1 INTRODUCTION

Dense retrieval models, such as ColBERT [24], ANCE [41], and DPR [23], have been adapted to cross language ad-hoc retrieval (CLIR) where queries and documents are in different languages by replacing monolingual embedding with a multilingual embeddings (e.g., mBERT [11] and XLM-R [8]). These dense retrieval models learn to encode queries and documents separately into fixed-length dense representations by fine-tuning a pretrained model (e.g, BERT [11]) with a retrieval objective using a large number of query-document pairs such as the ones from MS MARCO [2] or Natural Questions [25]. Recent work showed that these models are effective for CLIR when trained with monolingual query-document pairs, enabling zero-shot transfer [28, 32, 37]. Alternatively, training the model with translated MS MARCO (translate-train) is more effective but also much more expensive [32, 39].

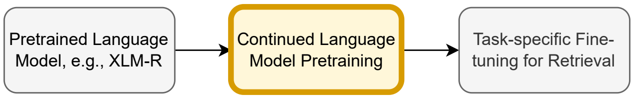
Pretrained Language Model, e.g., XLM-R
Continued Language Model Pretraining
Task-specific Fine-tuning for Retrieval

Figure 1: Pipeline for training a dense retrieval model. We propose an additional pretraining phase targeting CLIR.

However, most pretrained language models do not explicitly ensure the representations of a pair of related texts are similar [15]. This calls for a task-specific fine-tuning process to retrofit the representation produced by the pretrained model to be closer between related or relevant text. Such processes can be complex and computationally expensive, such as RocketQA [35], and, thus, efficient multi-stage training that introduces a "continued pretraining" to the pipeline was proposed for monolingual retrieval [16, 20] before running a task-specific fine-tuning with retrieval objectives (illustrated in Figure 1).

By construction, the representations for the text conveying similar information in different languages are not necessarily similar, since multilingual pretrained models such as mBERT and XLM-R do not introduce parallel text during pretraining. In other settings, incorporating alignment information into the retrieval model has been shown to be useful for CLIR [10, 19]. We hypothesize that explicitly promoting token-level similarity during the pretraining phase will enhance the effectiveness of CLIR models.

To address the aforementioned issues, we propose C3, a continued pretraining approach leveraging weak supervision with document-aligned comparable corpora to encourage representations of the text with similar meaning in different languages to be more similar using contrastive learning. This continued pretraining phase modifies an off-the-shelf pretrained model before it is fine-tuned to the actual retrieval objective, as illustrated in Figure 1. Specifically, we model the similarity between a pair of texts using contrastive learning with token-level embeddings to encourage the

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531886

model to embed token-level similarity and alignment information. We use Wikipedia articles in the language pair of interest, linking them based on cross-language links present in Wikipedia. We test this using high-resource languages for which strong evaluation resources are available. but this weakly supervised approach could also be applied to lower resource languages in which alternative approaches that rely on parallel text might prove impractical.

To summarize, our contributions are two-fold. First, we propose to continually pretrain the model for CLIR dense retrieval to promote similar representations between texts with similar meanings in different languages, using contrastive learning. To our best knowledge, this is the first work that applies contrastive learning to CLIR pretraining in this way. Secondly, we do this in a way that relies only on weakly supervised document-scale links in Wikipedia.

2 BACKGROUND AND RELATED WORK

Since the introduction of pretrained transformer-based language models, neural retrieval models have been taking advantage of these models for more effective query-document matching. Early work in monolingual retrieval involved building cross-encoder models [9, 29, 33] that leveraged the full interaction between the queries and documents to produce the relevance scores. Subsequently, similar models [19, 22, 38, 42, 44] were adapted to the CLIR setting. While effective, such models can only rerank documents since they process both queries and documents during inference and thus yield a longer running time compared to traditional sparse retrieval techniques such as BM25 [36]. DPR-style dense retrieval models overcome this limitation by scoring the documents based on the similarity of their representations, which allows the language models to encode documents beforehand.

However, the representations produced by off-the-shelf language models are undertrained [16] and thus directly scoring documents with such representations yields suboptimal retrieval results [27]. Additional task-specific fine-tuning with relevance labels produces much better representations on either the sequence level, such as ANCE [41] and DPR [23], or the token level, such as ColBERT [24]. Especially for sequence level representations, often summarized in the CLS token, contrastive learning [6, 34] that trains a model with one positive and multiple negative examples for each query has been shown to be one of the most effective training techniques for dense retrievers [21, 35, 41]. In-batch negative sampling further reduces the memory requirement by treating positive examples for other queries as negative [16, 23]. A similar objective was utilized to pretrain the cross-encoder retrieval model for the CLIR task [43].

Continuing pretraining the off-the-shelf language model has been investigated in mono-lingual retrival [5, 13, 16]. Specifically, coCondenser [16] continued pretraining of the language model with a passage-containing classification task (i.e., determining if a pair of passages belong to the same document) through contrastive learning on the representation of the passages for monolingual IR before fine-tuning it as a DPR model. This weakly supervised pretraining teaches the model to bring the representation of passages extracted from the same documents closer, which benefits the downstream dense retrieval task by assuming passages from the same document convey a similar meaning. coCondenser also trains with a masked language modeling task on the Condenser head [15] that adds two additional layers at the end of the network with the embeddings of CLS from the last layer and the rest from the middle layer. This Condenser head is removed after pretraining, but it has been shown to adjust the encoder effectively.

Recently, dense retrieval models have been adapted to CLIR by replacing the encoder with a multilingual pretrained model, such as mBERT, XLM or XLM-R [3, 27, 32]. To utilize existing monolingual collections with a large number of relevance labels such as MS MARCO [2], dense retrievers with multilingual embeddings can be trained on such corpora with zero-shot transfer to CLIR by leveraging the multilinguality of the encoder [28, 32]. Alternatively, with the help of translation models, one can translate the monolingual training collection into the language pair of interest and train the retriever on it (a "translate-train" approach) [32, 38]. This training scheme encourages the model to bring the representations of related queries and documents closer across languages. However, training effective translation models can be resource-intensive.

Besides the challenges in obtaining the translation, teaching the models two complex tasks jointly can also be tricky. Learning from coCondenser, a two-stage process with a continued language model pretraining followed by task-specific fine-tuning can help the model acquire knowledge incrementally. The following section introduces a pretraining approach that encourages the model to bring the representations of passages in different languages with similar meanings closer before fine-tuning with retrieval objectives.

3 C3: CONTINUED PRETRAINING WITH CONTRASTIVE LEARNING FOR CLIR

In this section, we introduce C3, a continued pretraining approach with contrastive learning that encourages similar representations of a pair of texts across languages. The language model learns to establish a semantic space containing the two languages of interest with meaningful similarity by training with this objective.

Specifically, consider a comparable corpus with linked document pairs $(d_i^{\mathcal{S}}, d_i^{\mathcal{T}})$ in languages $\mathcal{S}$ and $\mathcal{T}$ (i.e., pairs of documents in different languages containing similar information). Given a list of such document pairs $[(d_1^{\mathcal{S}}, d_1^{\mathcal{T}}), (d_2^{\mathcal{S}}, d_2^{\mathcal{T}}), \ldots, (d_n^{\mathcal{S}}, d_n^{\mathcal{T}})]$, we construct a list of spans $[s_1^{\mathcal{S}}, s_1^{\mathcal{T}}, s_2^{\mathcal{S}}, s_2^{\mathcal{T}}, \ldots, s_n^{\mathcal{S}}, s_n^{\mathcal{T}}]$ by randomly sampling one span from each document.

Let h_i^L be the sequence of token representations of span s_i^L where $L \in \{\mathcal{S}, \mathcal{T}\}$, we construct its SimCLR [6] contrastive loss as

$$\mathcal{L}_{iL}^{co} = -\log \frac{\exp\left(f\left(h_i^{\mathcal{S}}, h_i^{\mathcal{T}}\right)\right)}{\sum_{j=1}^{n} \sum_{k \in \{\mathcal{S},\mathcal{T}\}} \mathbb{1}(i \neq j \wedge L \neq k) \exp\left(f\left(h_i^{l}, h_l^{k}\right)\right)}$$

with $\mathbb{1}(\bullet)$ being the indicator function and $f(h_1, h_2)$ being the similarity function between representations h_1 and h_2. This contrastive loss is similar to the one proposed in coCondenser [16] but encourages the model to learn different knowledge. Instead of sampling pairs of spans from the same document, we construct the pair by sampling one span from each side of the linked documents. Equation 3 promotes the representation $h_i^{\mathcal{S}}$ and $h_i^{\mathcal{T}}$ to be closer while discouraging representations of spans in the same language from being similar (since k can the same as L). This construction pushes the encoder away from clustering text in the same language in

Table 1: Reranking effectivness of ColBERT and DPR models with and without our C3 pretraining. The top shows XLM-RoBERTa-base models; the bottom shows XLM-algin-base models. Symbols indicate statistically significant differences at $p < 0.05$ by a two-tailed paired t-test with Bonferroni correction for 6 tests, either with and without C3 (*) or between C3 and original BM25 results (†). Δ shows the mean relative improvement from C3 across the 6 collections.

		nDCG@100							nDCG@10						
Retrieval	With	HC4		NTCIR	CLEF				HC4		NTCIR	CLEF			
Model	C3	Chinese	Persian	Chinese	Persian	German	French	Δ	Chinese	Persian	Chinese	Persian	German	French	Δ
QT + BM25		0.362	0.354	0.264	0.336	0.419	0.563		0.258	0.251	0.229	0.407	0.379	0.505	
XLM-RoBERTa (base)															
ColBERT	✗	0.352	0.385	0.249	0.283	0.510	0.590		0.248	0.277	0.223	0.325	0.513	0.514	
	✓	***0.444**	**0.391**	**0.278**	†***0.286**	†**0.521**	0.574	+8%	***0.345**	0.274	**0.255**	**0.337**	†**0.535**	0.482	+11%
DPR	✗	0.330	0.319	0.218	0.259	0.467	0.531		0.223	0.220	0.184	0.299	0.449	0.449	
	✓	***0.395**	**0.341**	**0.255**	†**0.266**	†**0.503**	**0.562**	+10%	***0.287**	**0.226**	**0.231**	†**0.302**	†***0.523**	**0.491**	+15%
XLM-align (base)															
ColBERT	✗	0.425	0.399	0.303	0.252	0.523	0.579		0.332	0.294	0.283	0.285	0.532	0.478	
	✓	†***0.483**	**0.400**	†**0.330**	**0.275**	†**0.528**	**0.588**	+4%	†***0.408**	0.280	†**0.316**	**0.321**	†**0.536**	**0.499**	+6%
DPR	✗	0.385	0.366	0.260	0.235	0.480	0.581		0.300	0.256	0.239	0.265	0.482	0.503	
	✓	**0.421**	**0.403**	**0.286**	†**0.244**	†**0.503**	**0.586**	+6%	**0.324**	**0.312**	**0.264**	†**0.279**	†**0.520**	**0.506**	+8%

the semantic space and pulls the text across languages with similar meanings closer, while retaining distributional robustness by randomly matching the spans in the documents.

To promote token-level similarities, we apply the MaxSim operator proposed in ColBERT [24] as the similarity function $f(h_1, h_2)$. Specifically, the function can be written as

$$f(h_1, h_2) = \sum_{i \in |h_1|} \max_{j \in |h_2|} h_{1i} \cdot h_{2j}^T$$

where $|h_\bullet|$ denotes the number of tokens in the corresponding span and $h_{\bullet k}$ denotes the representation of the k-th token in $h_\bullet$. With this similarity function, the contrastive learning loss flows into the token representation to explicitly promote token alignment in the semantic space.

Finally, we combine $\mathcal{L}_{iL}^{co}$ with the masked language modeling loss $\mathcal{L}_{iL}^{mlm}$ and $\mathcal{L}_{iL}^{cdmlm}$ on span s_i^L from the transformer network and the Condenser head [15], respectively, to train the bottom half of the network more directly. Therefore, the total loss $\mathcal{L}$ can be expressed as

$$\mathcal{L} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{L \in \{\mathcal{S}, \mathcal{T}\}} \left[\mathcal{L}_{iL}^{co} + \mathcal{L}_{iL}^{cdmlm} + \mathcal{L}_{iL}^{mlm} \right]$$

4 EXPERIMENTS AND ANALYSIS

Our experiment follows the workflow in Figure 1. In this specific study, we use English as our pivot language for the queries and Chinese, Persian, French, and German as our document languages. However, we argue that C3 is generalizable to other language pairs. In the rest of the section, we discuss our experiments' data, models, setup, and results.

4.1 Datasets

To continue pretraining the off-the-shelf pretrained models with C3, we assembled linked Wikipedia articles on the same topic in different languages. Specifically, we leveraged CLIRMatrix [40], a retrieval collection that uses the article titles as the queries to retrieve documents for 19,182 language pairs. For each language pair, we extract all query and document pairs with relevance score 6, which are the Wikipedia pages on the same topic as asserted by inter-wiki links (one query only has one document with a score of 6 given a specific language pair). These documents are linked to construct the comparable corpus. We extracted 586k, 586k, 1,283k, and 1,162k document pairs for Chinese, Persian, French, and German, respectively.

For task-specific fine-tuning, we use the "small" training triples provided in MSMARCO-v1, which consists of 39 million triples of query, positive, and negative passages.

We evaluate the final retrieval models on HC4 [26], a newly constructed evaluation collection for CLIR, for Chinese and Persian, NTCIR [31] for Chinese, CLEF 08-09 for Persian [1, 14], and CLEF 03 [4] for French and German. HC4 consists of 50 topics for each language. We have 100 topics for NTCIR and CLEF 08-09 Persian and 60 topics for CLEF 03 French and German. We use the title in English as our evaluation queries.

Despite experimenting with relatively high resource language pairs, we argue that there is no language-specific component in C3. We believe C3 is applicable to language pairs that have similar amount of linked Wikipedia pages.

4.2 Experiment Setup

We test our proposed approach with XLM-R-base [8]. Additionally, we also tested XLM-align-base [7], which is a variant of XLM-R-base pretrained with parallel text in 14 language pairs and multilingual text in 94 languages. All text in our experiments is tokenized by Sentence BPE [8], which XLM-R uses.

We construct the spans from document pairs with a window of 180 tokens. We pretrain the model with C3 for 100,000 gradient update steps with an initial learning rate set to 5×10^{-6} using 4 GPUs with 24GB of memory each. We leveraged Gradient Cache [18] to run with batches of 64 document pairs (16 per GPU).

Table 2: Ablation study on different similarity function used in contrastive learning with and without the Condenser head (Cond.). The values showed in the table is nDCG@100 on HC4 Chinese test set.

			Contrastive Similarity		
Lang. Model	Ret. Model	Cond.	None	CLS	MaxSim
XLM-R	ColBERT	✗	0.352	0.389	0.410
		✓	–	0.391	0.444
	DPR	✗	0.330	0.382	0.381
		✓	–	0.368	0.395
XLM-A	ColBERT	✗	0.425	0.482	0.474
		✓	–	0.457	0.483
	DPR	✗	0.385	0.406	0.406
		✓	–	0.408	0.421

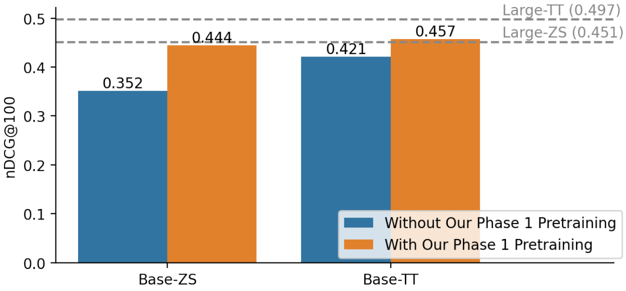

Figure 2: ColBERT models with zero-shot transfer (ZS) and translate-train (TT) approaches using XLM-RoBERTa-base on HC4 Chinese test set. The dashed line demonstrate the nDCG@100 value for XLM-RoBERTa-large with both approaches.

We tested on two dense retrieval models: ColBERT [24] and DPR [23]. After pretraining, each model is fine-tuned with the retrieval objective (either ColBERT or DPR) for 200,000 steps also using 4 GPUs with a learning rate set to 5×10^{-6} for each query-document language pair. We use the original implementation of ColBERT for its fine-tuning and Tevatron [17] for DPR with a shared encoder for queries and documents. Both retrieval models are tested in a reranking setting, where all models rerank the top-1000 documents retrieved by BM25 with machine translated queries. The machine translation models for Chinese and Persian were trained using AWS Sockeye v2 Model [12] with 85M and 12M general domain parallel sentences for each language pair respectively. We used Google Translate for German and French.

4.3 Results and Analysis

Table 1 summarizes the main results of our experiments, which indicate that building dense retrieval models using C3 yields better effectiveness. When starting from XLM-R, C3 provides an 8% relative improvement in nDCG@100 (and 11% in nDCG@10) over directly fine-tuning a ColBERT model. The model benefits from the warm start before training with relevance labels by pretraining with a similar objective (MaxSim) with weakly supervised text. On the other hand, we observe a slightly larger gain on DPR, suggesting even retrieval models that score documents with sequence representations (i.e., embeddings of CLS tokens) benefit from a task that promotes token-level similarity.

The improvement in the retrieval effectiveness by C3 is less substantial when starting from XLM-align (at most 6% in nDCG@100 compared to 10%). Since XLM-align is trained with parallel text, its ability to create relationships between text across languages is better than XLM-R, resulting in a diminishing return from investing computation resources in pretraining. Nevertheless, C3 still provides more effective retrieval results across languages.

Among the evaluated language pairs, English-French is particularly interesting. Applying C3 yields negative "improvements" in some cases. As English and French have a close relationship linguistically, we suspect the original XLM-R model, which is not trained with parallel text, already establishes an effective cross-language semantic space. Continued pretraining with C3 may simply not be necessary in such a case. Notably, XLM-align, which initialized its parameters by XLM-R, also yields worse retrieval results (0.590 to 0.579 in nDCG@100 and 0.514 to 0.478 in nDCG@10), which further supports our observation.

Note that all our reranking models underperform BM25 on CLEF Persian collection. After evaluating separately on topics generated in CLEF 2008 and 2009, we discovered that the topic characteristics are different between the two (nDCG@100 of 0.421 on 08 and 0.250 on 09 for BM25). Models pretrained with C3 underperform BM25 in 2008 topics, but are at least on par with BM25 on 2009 topics. While this effect deserves further investigation, we note that queries for this collection were originally created in Persian and then translated into English, possibly initially by nonnative speakers [1, 14]. Perhaps the English queries in 2009 better match the English for which our models have been trained. Nevertheless, C3 still improves the pretrained language models in these cases.

Comparing the average relative improvements (over all six test collections) that result from applying C3, we consistently see somewhat stronger relative improvements with nDCG@10 than with nDCG@100. From this we conclude that the effects of the improved modeling are particularly helpful nearer to the top of the ranked list, where interactive users might be expected to concentrate their attention.

To investigate our hypothesis regarding the utility of token-level similarity, we evaluate models in which different similarity functions were used as a basis for contrastive learning in continued pretraining. Using the CLS token in this way is similar to the coCondenser model. Results in Table 2 suggest that with the Condenser head, as implemented in the coCondenser model, pretraining with MaxSim similarity as the contrastive learning objective produces better retrieval models. The improvement is minimal without the Condenser head, indicating that token-level similarity benefits from routing information directly to the bottom half of the network. Interestingly, the second-best approach among the four combinations is CLS-based contrastive learning without using the Condenser head, which contradicts the original proposal of coCondenser. However, any continued pretraining is rewarding. Despite the competition among the variants, all language models with continued pretraining outperform their original off-the-shelf version.

Finally, we ask the question: what if we can afford to translate MS MARCO so that we can use a translate-train model? To investigate, we utilize the Chinese translation of the MSMARCO-v1 training triples from ColBERT-X [32], which can also be accessed via `ir_datasets` [30] with the dataset key neumarco/zh[1]. Figure 2 shows that without C3, the ColBERT model improves from 0.352 to 0.421, which is still worse than zero-shot transfer models trained with C3 for CLIR, suggesting allocating effort to C3 rather than training a translation model when computational resources are limited. When both are affordable, the effectiveness (0.457) is on par with zero-shot transfer a ColBERT model with XLM-R-large (0.451), which is even more expensive to train. With translate-train, ColBERT with XLM-R-large achieves close to 0.5 in nDCG@100 but requires more computational resources to run.

5 CONCLUSION AND FUTURE WORK

This paper proposed a continued pretraining task C3 on a weakly supervised corpus with a contrastive learning objective. We showed that the final retrieval models that are fine-tuned from models trained with C3 are more effective than off-the-shelf multilingual models. Further analysis suggests that translate-train can further improve retrieval models fine-tuned from C3-pretrained models. Evaluating with larger models such as XLM-R-large can also provide insight into the robustness of our approach. A natural next step would be extending C3 on some lower-resource languages where we have fewer Wikipedia articles in such languages.

Beyond that, despite being motivated by CLIR problems, C3 might also be applied to monolingual retrieval in cases where we have documents on the same topic that may use different writing styles.

REFERENCES

[1] Eneko Agirre, Giorgio Maria Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. 2008. CLEF 2008: Ad hoc track overview. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 15–37.

[2] 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. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL]

[3] Hamed Bonab, Sheikh Muhammad Sarwar, and James Allan. 2020. 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*. Association for Computing Machinery, New York, NY, USA, 9–18.

[4] Martin Braschler and Carol Peters. 2004. CLEF 2003 Methodology and Metrics. In *Comparative Evaluation of Multilingual Information Access Systems*. Springer Berlin Heidelberg, 7–20.

[5] Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932* (2020).

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[7] Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. *arXiv preprint arXiv:2106.06381* (2021).

[8] 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. arXiv:1911.02116 [cs.CL]

[9] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 985–988.

[10] Kareem Darwish and Douglas W Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. 338–344.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The Sockeye 2 Neural Machine Translation Toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Association for Machine Translation in the Americas, Virtual, 110–115.

[13] Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang, Jiafeng Guo, and Yiqun Liu. 2021. Pre-training Methods in Information Retrieval. *arXiv preprint arXiv:2111.13853* (2021).

[14] Nicola Ferro and Carol Peters. 2009. CLEF 2009 ad hoc track overview: TEL and persian tasks. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 13–35.

[15] Luyu Gao and Jamie Callan. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. *arXiv preprint arXiv:2104.08253* (2021).

[16] Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540* (2021).

[17] Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. 2022. Tevatron: An Efficient and Flexible Toolkit for Dense Retrieval. *ArXiv* abs/2203.05765 (2022).

[18] Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. *arXiv preprint arXiv:2101.06983* (2021).

[19] Zhiqi Huang, Hamed Bonab, Sheikh Muhammad Sarwar, Razieh Rahimi, and James Allan. 2021. *Mixed Attention Transformer for Leveraging Word-Level Knowledge to Neural Cross-Lingual Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 760–770. https://doi.org/10.1145/3459637.3482452

[20] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards Unsupervised Dense Information Retrieval with Contrastive Learning. arXiv:2112.09118 [cs.IR]

[21] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv preprint arXiv:2112.09118* (2021).

[22] Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual Information Retrieval with BERT. (April 2020). arXiv:2004.13005 [cs.IR]

[23] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. arXiv:2004.04906 [cs.CL]

[24] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. arXiv:2004.12832 [cs.IR]

[25] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[26] Dawn Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. 2022. HC4: A New Suite of Test Collections for Ad Hoc CLIR. (2022). https://arxiv.org/abs/2201.09992

[27] Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. On Cross-Lingual Retrieval with Multilingual Text Encoders. *arXiv preprint arXiv:2112.11031* (2021).

[28] Sean MacAvaney, Luca Soldaini, and Nazli Goharian. 2020. Teaching a New Dog Old Tricks: Resurrecting Multilingual Retrieval Using Zero-shot Learning. In *Proceedings of the 42nd European Conference on Information Retrieval Research*. 246–254. https://doi.org/10.1007/978-3-030-45442-5_31

[29] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1101–1104.

[30] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified Data Wrangling with ir_datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2429–2436.

[1] https://ir-datasets.com/neumarco.html#neumarco/zh

[31] T MITAMURA. 2010. Overview of the NTCIR-8 ACLIA Tasks: Advanced cross-lingual information access. In *NTCIR-8 Workshop, 2010*.

[32] 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. (2022). https://arxiv.org/abs/2201.08471

[33] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-Stage Document Ranking with BERT. arXiv:1910.14424 [cs.IR]

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[35] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191* (2020).

[36] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *TREC*.

[37] Peng Shi and Jimmy Lin. 2019. Cross-lingual relevance transfer for document retrieval. *arXiv preprint arXiv:1911.02989* (2019).

[38] P Shi and J Lin. 2019. Cross-lingual relevance transfer for document retrieval. *arXiv preprint arXiv:1911.02989* (2019).

[39] Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-lingual training with dense retrieval for document retrieval. *arXiv preprint arXiv:2109.01628* (2021).

[40] Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4160–4170.

[41] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. arXiv:2007.00808 [cs.IR]

[42] P Yu and J Allan. 2020. A Study of Neural Matching Models for Cross-lingual IR. *Proceedings of the 43rd International ACM SIGIR* (2020).

[43] Puxuan Yu, Hongliang Fei, and Ping Li. 2021. Cross-lingual Language Model Pretraining for Retrieval. In *Proceedings of the Web Conference 2021*. 1029–1039.

[44] Lingjun Zhao, Rabih Zbib, Zhuolin Jiang, Damianos Karakos, and Zhongqiang Huang. 2019. Weakly supervised attentional model for low resource ad-hoc cross-lingual information retrieval. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. 259–264.