Cross-lingual Language Model Pretraining for Retrieval

Puxuan Yu, Hongliang Fei, Ping Li
Cognitive Computing Lab
Baidu Research
Bellevue, WA, USA
{pxyuwhu,feihongliang0,pingli98}@gmail.com

ABSTRACT

Existing research on cross-lingual retrieval cannot take good advantage of large-scale pretrained language models such as multilingual BERT and XLM. We hypothesize that the absence of cross-lingual passage-level relevance data for finetuning and the lack of query-document style pretraining are key factors of this issue. In this paper, we introduce two novel retrieval-oriented pretraining tasks to further pretrain cross-lingual language models for downstream retrieval tasks such as cross-lingual ad-hoc retrieval (CLIR) and cross-lingual question answering (CLQA). We construct distant supervision data from multilingual Wikipedia using section alignment to support retrieval-oriented language model pretraining. We also propose to directly finetune language models on part of the evaluation collection by making Transformers capable of accepting longer sequences. Experiments on multiple benchmark datasets show that our proposed model can significantly improve upon general multilingual language models in both the cross-lingual retrieval setting and the cross-lingual transfer setting.

ACM Reference Format:
Puxuan Yu, Hongliang Fei, Ping Li. 2021. Cross-lingual Language Model Pretraining for Retrieval. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442381.3449830

1 INTRODUCTION

Cross-lingual ad-hoc retrieval (CLIR) refers to the task of retrieving documents in the target language L_t with queries written in the source language L_s. A search engine with better CLIR capability has broader impact, as it can fulfill information needs of more users across language barriers.

Recently, the use of monolingual pretrained language models based on Transformer [46] neural networks (e.g., BERT [13]) for ad-hoc retrieval in English has advanced the performance in the literature to a large degree. For instance, almost all leading competitors in the MS MARCO[1] passage and document retrieval tasks rely on Transformer-based pretrained language models. In the meantime, multilingual language models (e.g., mBERT [13] and XLM [9]) were proposed, and they have been proven to perform well on various downstream cross-lingual tasks, such as cross-lingual text classification, cross-lingual named entity recognition, and supervised/unsupervised machine translation. Nevertheless, the wave of multilingual language models has not yet benefited CLIR.

Most related research focuses on the success of *cross-lingual relevance transfer* [32, 43]. The language model is first finetuned on monolingual collection in language L_s with more labelled data, and then applied for inference to *monolingual* retrieval in another language L_t, where there is usually less available training data. This task, though significantly important, is different from cross-lingual ad-hoc retrieval. In CLQA literature [29], cross-lingual relevance transfer is directly referred to as cross-lingual transfer (dubbed XLT), while the "real" cross-lingual task where question and context are in different languages is called *generalized* cross-lingual transfer (G-XLT). In this paper, we inherit this naming convention and focus on the G-XLT setting.

The state-of-the-art methodology for CLIR is generally using learning-to-rank with neural matching models [53] coupled with pre-acquired cross-lingual word embeddings (CLE) [4, 26]. A few endeavors to adopt multilingual language models for CLIR have shown that such models perform inferior to a big margin compared with learning-to-rank with CLE [4]. Obviously, there is a gap between how language models should be used for monolingual (English) ad-hoc retrieval and cross-lingual ad-hoc retrieval. This research focuses on closing this gap.

We first consider the differences between pretraining and applying cross-lingual LMs. The prerequisite assumption to use cross-lingual LM for retrieval is that representations are well aligned across languages on multiple levels of text segments (i.e., word, sentence, paragraph and document). Conneau et al. [10] showed that representations from *monolingual* BERT in different languages can be linearly mapped to one another on both word and sentence levels, and that the success of a unified cross-lingual LM is mostly due to parameter sharing in upper encoder layers. Both mBERT and XLM focus on word-level and sentence-level tasks during pretraining: the masked language modeling task (MLM) trains the model to fill the blanks of monolingual sentences, while the translation language modeling task (TLM) challenges the model to fill the blanks in pairs of parallel sentences. The fact that they perform well on word and sentence level tasks but poorly on retrieval tasks suggests that representations of longer sequences [25] might not be well aligned in cross-lingual LMs. To that end, we propose two novel pretraining objectives for better aligning representation of longer texts and better modeling of query-document interactions. The query language modeling task (QLM) masks some query tokens and asks the model to predict the masked tokens based on query contexts and full relevant *foreign* document. We specifically increase the masking probability compared to autoencoder language modeling tasks to enforce referencing cross-lingual long sequences. The relevance

[1] https://microsoft.github.io/msmarco/

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449830

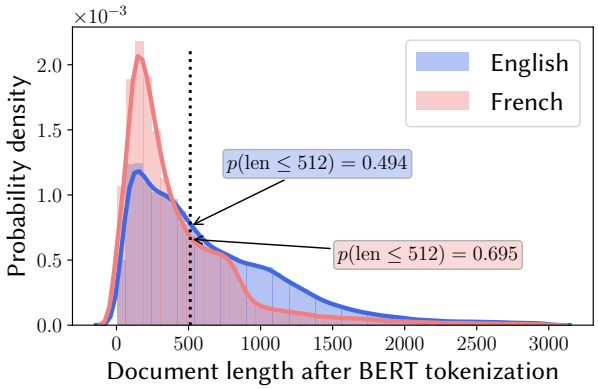

Figure 1: Distribution of document length in CLEF dataset. 512 tokens is marked for reference. Realistically, documents with 512 tokens cannot fit considering length of queries.

ranking task (RR) operates on more coarse-grained representations and directly resembles the cross-lingual ad-hoc retrieval task. Given a query and several *foreign* documents, the model is asked to rank these documents based on levels of relevance.

Our cross-lingual LMs are not fully "self-supervised", as we require some knowledge about query-document relevance for both pretraining objectives. We describe a simple yet effective approach for building such distant weakly-supervised data from multilingual Wikipedia (Wiki). Specifically, we match sections of multilingual versions of Wiki entities based on cross-lingual representation of section titles. For each pair of matched Wiki sections, we sample a sentence from one section as query, and use the other section as relevant document. Wiki sections are sources for more fine-grained semantics in various retrieval datasets [6, 14], and we extend the idea to multilingual Wiki [3, 12, 35] in this work. We end up with millions of raw matched sections for each language pair. Therefore, we consider this data extraction scheme to be a good balancing point between quantity and quality for our pretraining tasks.

We also re-evaluate the *pretrain-finetune-inference* paradigm of using language models for ad-hoc retrieval. Due to the square time and memory complexity of Transformer's full self-attention mechanism [46], Transformers and thus language models have a small upper limit on the input sequence length (e.g., 512 tokens for BERT). However, in most circumstances, 512 tokens is not enough to encode the query and the full document whilst performing finetuning (Figure 1). Current research on monolingual retrieval either truncate the documents such that the input sequences meet the size requirement [22, 36], or finetune language models on passage-level collections and then perform inference on longer test collection by post-aggregating relevance scores of document segments [52]. Truncating documents results in some degree of information loss. Also, there does not exist any multilingual passage-level relevance dataset like MS MARCO for English retrieval. We seek to finetune LMs for downstream retrieval tasks directly on evaluation collections (similar to non-retrieval cross-lingual tasks), but also minimizing information loss in the process. To that end, we replace the self-attention mechanism in Transformer with the global+sliding-window (GSW) attention [2] to unlock the ability of cross-lingual LM to process longer sequences in the "inside-Transformer" way. Note that there is also an "outside-Transformer" solution, where the original Transformer slides over a document, and a parameterized saturation function aggregates the windows and outputs a score [21]. In comparison, our solution is more computationally efficient, especially considering we also perform large-scale pretraining besides finetuning. We did not conduct comparison of these two methods in terms of effectiveness, which is left for future work.

The contributions of this paper can be summarized as follows:

- We propose two novel retrieval-oriented tasks for pretraining cross-lingual language models. We build weak-supervision data to support cross-lingual LM pretraining with our tasks.
- We employ the global+sliding-window attention in our cross-lingual language models to better align longer text representations across languages in both pretraining and finetuning stages, whilst minimizing information loss.
- We extensively evaluate our proposed models on downstream CLIR and CLQA tasks. We also conduct detailed experiments to support the rationale of each component from the empirical perspectives. For CLIR, we achieve 13.9% – 29.7% MAP improvement over vanilla mBERT re-ranker in all 12 language pairs on the bench-marking CLEF dataset. For CLQA, we see 1.7 and 2.8 point F1 improvement under XLT setting (German and Spanish), and 3.6 – 9.8 point F1 improvement under G-XLT setting (6 language pairs) over mBERT on the MLQA dataset.

Table 1: Frequently used acronyms in this paper.

CLE	Cross-lingual word embeddings.
(G-)XLT	(Generalized) Cross-lingual Transfer.
MLM	Masked Language Modeling task [13].
TLM	Translation Language Modeling task [9].
QLM	Query Language Modeling task proposed in this paper.
RR	Relevance Ranking modeling task proposed in this paper.
XLM(-R)	Cross-lingual language models proposed in [8, 9].
GSW	Global+Sliding Window attention mechanism [2].

2 RELATED WORK

2.1 Cross-lingual Ad-hoc Retrieval

Cross-lingual ad-hoc retrieval has always been considered as the combination of machine translation and monolingual ad-hoc retrieval. The initial translation resources are borrowed from the field of statistical machine translation (SMT). Some earlier works [20] use word-by-word translation. Ture and Lin [45] used translation tables from SMT to translate query into structured probabilistic structured query [11]. CLIR methods gradually shift towards using cross-lingual word embeddings [42] as translation resources. There are generally two ways to acquire CLE: pseudo-bilingual [4, 47] and post-projection [1]. Litschko et al. first proposed heuristics to use CLE for cross-lingual ad-hoc retrieval [30]. Most recently, the combination of neural matching models and CLE was proposed for document re-ranking and has yielded impressive performance on standard benchmarks [4, 53].

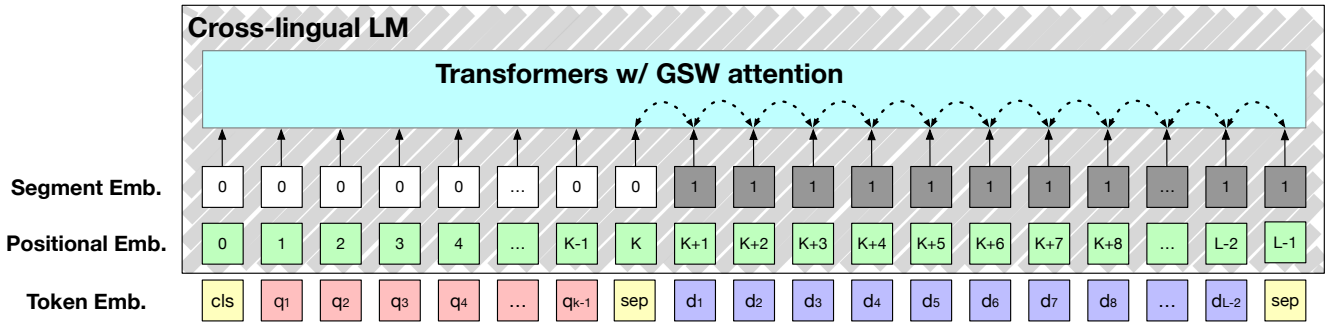

Figure 2: Our cross-lingual language model in detail. Document tokens can only attend to neighboring tokens in a w-token window. For demonstration, we let $w = 2$ in the figure. In experiments we use larger window sizes ($w = \{32, 64, 128, 256\}$).

2.2 Pretrained LM for Monolingual IR

Pretrained language models [13, 31, 39, 50] have brought a revolution to the field of human language technologies in general. We use BERT as an example here. There are two main approaches to apply BERT for ad-hoc retrieval: (i) **single-tower**: a query-document pair is packed into one sequence, separated by an `[sep]` token and then fed into one BERT encoder. Every query/document token can attend to the whole sequence during encoding (also referred to as cross-attention [6]). We take the output representation of the `[cls]` token for predicting ranking score; and (ii) **two-tower**: query and document are encoded with separate BERT encoders. The matching score is the cosine similarity of the two sequence embeddings. Two-tower models are more efficient for indexing query and document vector representations and are usually used in first-stage retrieval [6], while single tower models with full cross-attention are usually used for final-stage document re-ranking [36, 49, 51, 52]. For document re-ranking, it was shown that incorporating term-level matching signals from contextualized word embeddings in addition to output `[cls]` vector from cross-attention can provide additional improvement [33]. We study the re-ranking problem in this work, and we use the single-tower model without resorting to term-level matching for simplicity.

2.3 Cross-lingual Pretrained LM

Cross-lingual pretrained language models [8, 9, 13, 23] are capable of simultaneously encoding texts from multiple languages. Multilingual BERT [13] takes the same model structure and training objective as BERT, but was pretrained on more than 100 languages on Wikipedia. In addition to the masked language modeling (MLM) objective, the XLM model [9] is also pretrained with the translation language modeling objective (TLM) to take advantage of parallel sentence resources if available: a pair of parallel sentences are randomly masked, and the language model is challenged to predict the masked tokens by attending to local contexts as well as distant *foreign* contexts. XLM-RoBERTa [8] improves upon XLM by incorporating more training data. Two additional word and sentence level tasks were proposed to pretrain the Unicoder [23]. Evaluations on a series of word-level and sentence-level *cross-lingual transfer* tasks have shown that these cross-lingual LMs have significant utilities for transferring language knowledge from high-resource languages to low-resource languages.

In the contexts of retrieval, there are also research works on cross-lingual transfer for ad-hoc retrieval [32, 43] and question answering [23, 29, 41]. But different from cross-lingual transfer, using single-tower model for cross-lingual retrieval requires the language model to encode two sequences (CLIR: query/document, CLQA: question/context) from different languages in the same pass. Bonab et al. [4] reported unsuccessful attempts to use single-tower model and CEDR-like [33] matching model for CLIR by stating "pre-trained models with many languages are not providing high gain for CLIR and needs further investigations for fine-tuning or training". To the best of our knowledge, there is only one detailed report about using single-tower model for CLIR [24], in which the proposed method decouples query into terms and document into sentences. Therefore, their model's complexity is squared on the basis of vanilla BERT cross-attention and thus far from practical. In comparison, our proposed model is capable of encoding bilingual full query and document in one pass.

3 MODEL STRUCTURE

The structure of our cross-lingual LM is illustrated in Figure 2. The input is always a packed sequence containing one query and one document with `[sep]` token in between. To encourage learning language-agnostic representations, unlike XLM, we do not supplement language-specific embeddings. Instead, we simply input segment embeddings to let the model differentiate between two parts of input.

A large portion of documents in CLIR datasets (and in real applications of CLIR generally) exceed the input length limit of mBERT and XLM (Figure 1). We seek to build a cross-lingual LM that can encode more document content at pretraining, finetuning and inference stage. Therefore, we adopt the attention mechanism proposed in Longformer [2] to replace full self-attention as with mBERT, such that each Transformer block can encode longer sequences. Compared with Hofstätter et. al's solution that slides vanilla Transformer over long documents for finetuning monolingual language models [21], our model is more computationally efficient especially at pretraining stage.

Longformer is the long-document Transformer, where the $O(n^2)$ complexity self-attention is replaced with a series of linear attention mechanisms [2]. Specifically, we adopt the "global+sliding window" (GSW) attention. We let query tokens have global attention and

limit document tokens to sliding-window attention. Within each Transformer, all query tokens can still attend to any other tokens in the sequence, but document tokens can only attend to tokens within a w-token wide window. Note that special tokens like `[sep]` and `[cls]` also have global attentions. The original Transformer [46] computes attention scores as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

GSW, on the other hand, uses two sets of projections, $\{Q_s, K_s, V_s\}$ to compute attention scores of sliding window attention, and $\{Q_g, K_g, V_g\}$ to compute attention scores for the global attention. Intuitively, we regard GSW as relocation of computation power: instead of letting two long-distance document tokens attend to each other, we can instead let query tokens attend to more document tokens. In practice, comparing (a) full self-attention with maximum sequence length 512 and (b) GSW attention with maximum sequence length 1024 and window size 64 both using single-tower model, we observe that (i) they consume similar GPU memory; (ii) GSW runs slightly slower; and (iii) GSW is capable of encoding documents *more* than twice as long. We perform empirical comparisons of their effectiveness in § 5.

4 RETRIEVAL-ORIENTED CROSS-LINGUAL LANGUAGE MODEL PRETRAINING

The goal of task-specific language model pretraining is to further enhance the model's performance on downstream tasks by taking advantage of weak-supervision data applied to task-specific modeling objectives [5, 7, 16, 19, 44]. We describe weak-supervision data construction, and retrieval-oriented cross-lingual modeling tasks in detail in this section.

4.1 Data

General requirements of the ideal pretraining data for our tasks include: (i) each positive example contains a pair of short text (query) and long text (document) in different languages; (ii) query and document are semantically related; (iii) the number of training examples is large enough. There is a trade-off between the granularity of semantic relatedness and the number of available training examples, and we think Wiki *sections* is a good balance point. To that end, we choose to first match multilingual Wiki sections, and then sample a sentence from one matched section as query, and use the other section as document. We are motivated by the discussion on the granularities of semantics in monolingual Wiki [6]. Our approach is conceptually similar to the Inverse Cloze Task (ICT), where one sentence is sampled from a Wiki paragraph as query, and the rest of the paragraph is treated as document. The key differences are that: (i) we expand from monolingual Wiki to multilingual Wiki; (ii) we keep longer texts (section v.s. paragraph) as document, which is more similar to downstream retrieval tasks.

However, accurate cross-lingual Wiki section alignment information is unavailable. Multilingual Wiki pages of the same entity are usually not translation of one another, and they are often organized to have different structures. In fact, section alignment of multilingual Wiki is itself an active research question [37]. We adopt an easy, efficient and yet effective method for section alignment based on cross-lingual word embeddings (CLE). Suppose Page_s and Page_t are two Wiki pages in source and target language respectively of the same entity. We define a section's title as its *immediate* preceding title. For each section $\text{Sec}_s(i)$ in Page_s, we acquire its title embedding by averaging the CLE of all its title's terms (with stopwords removed), and similarly for each section $\text{Sec}_t(j)$ in Page_t. If the cosine similarity of title embeddings of $\text{Sec}_s(i)$ and $\text{Sec}_t(j)$ is greater than a threshold value η, we consider them as matched sections. The underlying assumptions are that: (i) titles are accurate summaries of section content; (ii) matched sections are related to the same aspect of the same entity. Conceptually, the relatedness of matched sections is "lower-bounded" such that in worst cases, two sections are related to different aspects of the same entity, which is still acceptable for retrieval [6]. Note that given two languages, we only allow one section to match with at most one foreign section, and we take the highest matching pair if there is conflict. The quality of the data can be reflected by performance of the pretrained model on downstream tasks in § 5.

Our model and data construction method support any language present in multilingual Wiki. We select four languages {English, Spanish, French, German}[2] for demonstration and convenience of evaluation. We use fastText CLE[3] and set $\eta = 0.3$[4]. We apply filter such that a pair of matched sections must have at least five sentences in both sections. The numbers of aligned sections we end up with are listed in Table 2.

Table 2: Number of aligned sections in each language pair.

En&De	En&Es	En&Fr	Es&De	Fr&De	Fr&Es
250.6K	295.8K	202.8K	169.9K	216.7K	171.4K

4.2 Pretraining Tasks

Given the described massive cross-lingual query-document relevance data, we introduce two novel pretraining tasks (Figure 3) for cross-lingual retrieval.

4.2.1 Query language modeling (QLM). Given a pair of cross-lingual query and document, we mask some percentage of query tokens at random, and then predict those masked tokens. The final hidden vectors corresponding to the mask query tokens are fed into an output softmax over the vocabulary. This task is conceptually motivated by the query likelihood model [38] where a query is assumed to be generated based on words that appear in a prototype document. If we mask 15% of query tokens as in [13], QLM becomes easier than MLM, because there is an extra full foreign document to support predictions. To that end, we increase the masking probability to 30% to enforce attention from query to foreign document. If we also mask document tokens, QLM reduces to approximate TLM [9]. However, masking document does not promote cross-lingual attention, as the information from short foreign query is neglectable for helping complete long document. Therefore, we think masking just query tokens with higher probability best promotes cross-lingual

[2]Language codes: En=English, Es=Spanish, Fr=French, De=German
[3]https://fasttext.cc/docs/en/aligned-vectors.html
[4]CLE usually has lower cosine similarities than monolingual embeddings [53]

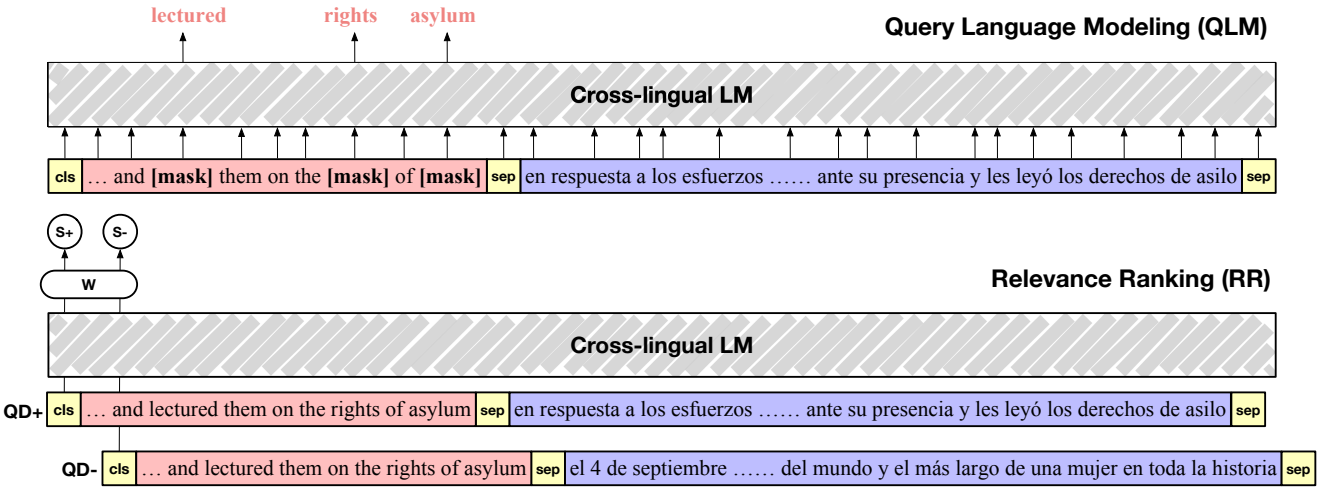

Figure 3: Proposed cross-lingual retrieval-oriented pretraining tasks.

query-document interaction understanding. We justify our choices using empirical experiments in § 5.3.2.

4.2.2 Relevance ranking (RR). This task directly resembles cross-lingual ad-hoc retrieval task, but the data has more coarse-grained semantics compared with finetuning data. Given a pair of cross-lingual query and document, we encode the packed sequence QD^+ with the model, and a learnable weight matrix $\mathbf{W}$ multiplies the output hidden vector of `[cls]` token and yield a ranking score S^+. Then, we randomly sample an irrelevant document, form sequence QD^-, and similarly acquire a ranking score S^-. We optimize the model with cross entropy loss, which supports multiple negative examples. We sample one negative per positive example in this work. Given query Q and document D, where $Q \in \text{Sec}_s(i) \in \text{Page}_s$ and $D = \text{Sec}_t(j) \in \text{Page}_t$, then $\{\text{Sec}_t(k)\}_{k \neq j}$ are considered hard negative examples. To avoid repeating the same negative document across training epochs when there are few sections in Page_t, the probability of sampling hard negative examples is dynamically adjusted according to the source entity, from which the positive query-document pair is created. The probability of sampling hard examples is set to $\alpha_T = 1 - (3/4)^T$, where T is the number of sections in Page_t. With probability $1 - \alpha_T$, we randomly sample a section in language L_t as irrelevant document.

4.3 Pretraining Details

Pretraining a language model from scratch is of high computational cost. We continue pretraining our retrieval-oriented language models from the public mBERT checkpoint[5]. Therefore, our cross-lingual LM is implicitly pretrained with three objectives (MLM, QLM and RR). We name different variants of our models as "(model, objective, epochs)". Model with full self-attention is directly called mBERT, while the model with GSW attention is named mBERT-GSW. For example, (mBERT-GSW, QLM-RR, 10 epochs) is the GSW attention model pretrained with both QLM and RR for 10 epochs. When pretraining with both objectives, we first train with RR in random order of language pairs, then train with QLM in random order of language pairs in each iteration. Each epoch contains 32K positive query-document pairs per language pair for each objective. We train our models with 16 Nvidia VOLTA (16GB) GPUs. We use Adam for model optimization [27]. Learning rate is set to 1e-5 and batch size is set to 32. We train for 20 epochs at maximum. Training mBERT takes about 24 hours, and training mBERT-GSW (window size=64, max sequence length=1024) takes about 40 hours.

We acquire mBERT-GSW with 1024 tokens by replacing the attention module in mBERT with GSW attention, and copying the positional embeddings from the first 512-token positions to the second 512-token positions [2]. Therefore, a naive mBERT-GSW is of worse language modeling ability (reflected in high bits-per-character measure) due to the copied non-optimized positional embeddings. Following [2], we use the MLM task on the Wikitext-103 dataset[6] and perform 2K gradient updates such that mBERT-GSW has similar BPC compared with base mBERT.

5 EXPERIMENTS

5.1 Cross-lingual Ad-hoc Retrieval

5.1.1 Evaluation Data and Metric. We adopt the standard gold CLIR dataset **CLEF** for evaluating cross-lingual ad-hoc retrieval effectiveness. We use the test collections from the 2000-2003 ad-hoc retrieval test suite[7] combined together. Two hundred topics in different languages are regarded as queries for retrieving news articles in different languages. We choose four languages in our pretraining data and thus form twelve cross-lingual query-document evaluation pairs. Following the standard practice [4, 30, 47], queries were created by concatenating the title and the description of each topic. Queries without any relevant document are removed. We do not employ a first-stage retrieval model like BM25 to get top candidates for re-ranking, for that such a method would require a query translation module and thus might introduce bias. Since human evaluators were presented with top-ranked documents when

[5]https://huggingface.co/bert-base-multilingual-uncased

[6]https://s3.amazonaws.com/research.metamind.io/wikitext/wikitext-103-raw-v1.zip

[7]http://catalog.elra.info/en-us/repository/browse/ELRA-E0008/

creating the relevance labels, we directly use all labelled documents with respect to a query as re-ranking candidates, following [53]. We report mean-average-precision (MAP) on query level. Statistically significant differences in MAP are determined using the two-tailed paired t-test with $p < 0.05$. The statistics of the CLEF dataset are shown in Table 3.

Table 3: Statistics of CLEF: number of queries (#query), average number of relevant (#pos) and irrelevant (#neg) documents per query, and average number of document tokens after mBERT tokenization (doc. length) for each language.

	#query	#pos	#neg	doc. length
En	176	18.5	404.0	699.2
Es	156	50.7	319.5	490.2
Fr	185	22.0	267.3	454.3
De	192	25.0	324.3	448.7

5.1.2 Competing Methods. We compare our models with several recent competitive CLIR methods.

(i) **BWE-AGG**. It is an unsupervised approach that first builds query and document embeddings by summing the CLE of their constituent terms [30]. Candidate documents are ranked by the cosine similarity of their embeddings with the query embeddings. There are two variants based on different summing weights for constructing document embedding: BWE-AGG-ADD uses uniform weight for all terms, and BWE-ADD-IDF weights document terms with IDF in the target language collection. We use fastText embeddings[3].

(ii) **TbT-QT-QL**. It is an unsupervised *query translation approach* based on CLE [30]. Each source language query term is translated to its nearest target language term in the CLE space. The CLIR task is thus reduced to monolingual retrieval task, and the translated queries are used with query likelihood model [38]. We use Galago[8] for building inverted indexes and retrieving documents. We use fastText embeddings[3] for query translation.

(iii) **DRMM and K-NRM**. We select the two matching models [17, 48] from an earlier study on neural CLIR [53]. They build term-level query-document interactions from CLE, but use different pooling methods to output matching scores. We implement the two models based on Matchzoo [18]. For CLE, we test fastText embeddings[3] and smart-shuffling bilingual word embeddings [4]. The former is an example of post-projection CLEs, and aligns fastText embeddings trained on monolingual Wikipedias in 44 languages into one space using the relaxed CSLS method [26]. Smart-shuffling is a pseudo-bilingual method, but instead of randomly shuffling words in parallel sentences, it also leverages word-level parallel data (i.e., translation dictionaries) to guide to process in order to bridge the "translation gap". We use smart-shuffling embeddings with window size set to 10, kindly provided by the authors. Note that the smart-shuffling embeddings are *bilingual*, and only overlap with our evaluation language on {En&Fr, En&De}. Therefore, we can only report its performance on four query-document language pairs.

(iv) **mBERT**. We use the public checkpoint of multilingual BERT[5]. It was originally pretrained with the masked language modeling (MLM) and next sentence prediction (NSP) objectives [13] on the top 102 languages with the largest Wikipedia dumps.

(v) **XLM-R**. We use the public checkpoint[9] [8]. It was originally pretrained with the MLM objective on the CommonCrawl corpus in 100 languages.

5.1.3 Evaluation Details. As mentioned earlier, we use all labelled query-document pairs on CLEF as hard candidates, and report re-ranking MAP. For unsupervised methods, we test on all queries. For methods that require training, we adopt five-fold cross validation to overcome the small number of queries per language pair. Evaluation is performed separately in terms of language pairs. Specifically, each training (finetuning) epoch contains all positive query-document pairs. Each positive document is paired with one randomly sampled hard negative document, and we optimize with pairwise cross entropy loss. Maximum number of training (finetuning) epochs is set to 20. We record MAP on test set when the model yields best MAP on valid set. For DRMM, bin size is set to 30 and histogram mode is set to "log-count". For KNRM, we set the number of Gaussian kernels to 20 (plus another one for exact matching), and σ to 0.1. For finetuning Transformer-based models (mBERT, XLM and ours), we only finetune the last three encoder layers to avoid overfitting. Also, before finetuning we re-initialize ("reset") the parameters of the last three encoder layers for better stability [34].

5.1.4 Results. The overall results of all competing CLIR models on all evaluation language pairs are summarized in Table 4. We provide detailed analysis below.

Unsupervised approaches: BWE-AGG and TbT-QT-QL are unsupervised CLIR methods based on fastText CLE. In most cases, TbT-QT-QL is better than BWE-AGG by a large margin, which is consistent with findings reported in prior research [30, 53]. However, these two studies only perform experiments where English is the query language, while our experiments are more comprehensive. We found that on some occasions ({De-Es, De-Fr, Es-De, Fr-De}), BWE-AGG performs closely or even slightly better than TbT-QT-QL. The latter heavily relies on the quality of top-1 term translation. We suspect that the German embeddings are not aligned well with Spanish/French embeddings in a way that provides quality top-1 nearest-neighbor term translation.

Neural matching: DRMM and KNRM represent the category of neural matching. We see a big drop from the numbers reported in [53]. The main difference between their evaluation and ours is that they truncate documents to the first 500 terms, while we keep everything. We observe similar performance with their report if we employ the same truncation strategy: neural matching performs significantly better than unsupervised methods, but still worse than mBERT baseline in this paper. This suggests that KNRM and DRMM cannot handle long documents very well. We also present the first direct empirical comparison of smart-shuffling embeddings and fastText embeddings for retrieval. We observe that smart-shuffling has smaller vocabulary coverage in the CLEF collection, which can be a big factor to its inferior performance. We leave the qualitative comparison of these two CLEs for future work.

General language models: this category includes XLM-R, mBERT and mBERT-GSW. XLM-R's bad performance is consistent with

[8]https://www.lemurproject.org/galago.php

[9]https://huggingface.co/xlm-roberta-base

Table 4: CLIR performance on CLEF. Numbers are MAP. Best performance on each language pair is marked bold. $*$ indicates statistically significant improvement over mBERT (paired t-test, $p < 0.05$).

Query language	De			En			Es			Fr		
Document language	En	Es	Fr	De	Es	Fr	De	En	Fr	De	En	Es
BWE-AGG-ADD	0.126	0.280	0.194	0.186	0.247	0.133	0.238	0.118	0.190	0.234	0.116	0.273
BWE-AGG-IDF	0.146	0.325	0.235	0.197	0.258	0.147	0.238	0.132	0.221	0.224	0.125	0.293
TbT-QT-QL	0.361	0.344	0.229	0.263	0.395	0.310	0.244	0.396	0.300	0.215	0.327	0.357
DRMM (smart-shuffling)	0.143	-	-	0.226	-	0.133	-	-	-	-	0.156	-
DRMM (fastText)	0.195	0.304	0.164	0.238	0.313	0.187	0.264	0.192	0.143	0.225	0.211	0.284
KNRM (smart-shuffling)	0.199	-	-	0.274	-	0.194	-	-	-	-	0.167	-
KNRM (fastText)	0.227	0.337	0.263	0.285	0.372	0.250	0.310	0.187	0.240	0.298	0.180	0.345
XLM-R	0.124	0.233	0.178	0.269	0.307	0.222	0.211	0.149	0.212	0.199	0.148	0.293
mBERT	0.400	0.507	0.395	0.465	0.545	0.437	0.463	0.438	0.412	0.442	0.419	0.519
Ours (mBERT-GSW)	0.419*	0.511	0.425*	0.501*	0.564*	0.477*	0.471	0.448	0.434*	0.479*	0.442*	0.543*
Ours (mBERT, QLM-RR, 10)	0.472*	0.555*	0.476*	0.524*	0.577*	0.495*	0.524*	0.489*	0.499*	0.518*	0.476*	0.588*
Ours (mBERT, QLM-RR, 20)	0.472*	0.573*	0.476*	0.534*	0.590*	0.516*	0.526*	0.491*	0.504*	0.524*	0.497*	0.594*
Ours (mBERT-GSW, QLM-RR, 10)	0.501*	0.581*	0.491*	0.536*	0.614*	**0.527***	0.540*	0.530*	0.519*	0.536*	0.520*	0.602*
Ours (mBERT-GSW, QLM-RR, 20)	**0.519***	**0.601***	**0.506***	**0.545***	**0.621***	0.524*	**0.554***	**0.545***	**0.534***	**0.550***	**0.532***	**0.616***

Bonab et. al's findings [4]. We are first (to the best of our knowledge) to report decent performance of mBERT on CLIR. It seems counter-intuitive that XLM-R performs much worse than mBERT: they have similar model structure and pretraining objective (MLM), but XLM-R is trained with more data and is reported to outperform mBERT on various cross-lingual tasks. We conduct controlled experiments to exclude tokenizers and text casing as factors. We suspect that the failure of XLM-R for CLIR is due to the way pretraining data is fed to the model: unlike BERT, XLM-R takes in *streams* of tokens such that a sequence in a mini-batch can contain more than two consecutive sentences [9]. This may work well for word-level tasks like extractive QA (shown in § 5.2), but could cause confusion for tasks like CLIR which require representation alignment of long texts. Comparing mBERT and mBERT-GSW, we observe statistically significant improvement of the latter upon the former in most circumstances. This indicates that the benefit of accepting longer input sequence is not limited to the pretraining phase, but also at the finetuning and inference phase. Less information loss during finetuning can lead to significant difference at inference time.

Retrieval-oriented language models: this category describes cross-lingual LMs pretrained with QLM and RR objectives. By comparing (mBERT, QLM-RR) to base mBERT, we can see significant improvement on re-ranking effectiveness on all language pairs. This proves that (i) our proposed pretraining objectives are effective for downstream CLIR task; and (ii) the pretraining weak-supervision data constructed with section alignment from multilingual Wiki is of high quality and the learned knowledge is generalizable to non-Wiki collections. By comparing retrieval-oriented LMs with GSW attention (mBERT-GSW, QLM-RR) to ones with self-attention (mBERT, QLM-RR), we observe additional statistically significant improvement. It indicates that natively expanding Transformers's input length to encode more document content can provide further benefits for ad-hoc retrieval, which are additional to the benefits of retrieval-oriented pretraining. Our full model (mBERT-GSW, QLM-RR, 20) provides up to 29.7% MAP improvement on vanilla mBERT re-ranker.

5.2 Cross-lingual Question Answering

Cross-lingual extractive question answering is a word-level retrieval task, and it does not have explicit connection with either of our pretraining objectives. Therefore, it can better demonstrate the ability of generalization of our proposed language model pretraining strategies.

5.2.1 Evaluation Datasets. We use the MLQA dataset [29] for testing. There is no dedicated training data in MLQA, and following standard practice, we perform finetuning with SQuAD v2.0 training data[10] [40], and use the dev and test sets in MLQA for evaluation under two settings. (i) **zero-shot XLT**: dev/test sets are *monolingual* QA in a language different from finetuning language; and (ii) **G-XLT**: question and context/answer in dev/test sets are in different languages. Note that G-XLT may not be zero-shot in terms of language because either query or context language might be English. As there is no French data in MLQA, we end up with six language pairs for G-XLT and two languages (Spanish and German) for XLT. We measure F1 score and exact-match score, which are standard metrics for extractive QA. SQuAD and MLQA are much larger than CLEF, so we are able to finetune all encoder layers of language models. We report results on test set when the model yields best F1 score on dev set. We still conduct statistical significance test using two-tailed paired t-test with $p < 0.05$.

5.2.2 Results. The overall results of CLQA are summarized in Table 5. Comparing general language models, XLM-R performs slightly better than mBERT under XLT setting, but much worse under G-XLT setting. mBERT-GSW performs slightly better than mBERT, but the improvement is mostly not statistically significant. Our language models pretrained with QLM and RR yield statistically significantly improvement over mBERT when fully trained, and the improvement is more significant on G-XLT than XLT. This

[10] SQuAD v1.1 used in [29] is no longer publicly available. Therefore, numbers in two papers are not directly comparable.

Table 5: CLQA performance on MLQA. Numbers are F1/Exact-Match scores (%) in percentile format by convention. Best performance on each language pair is marked bold. * indicates statistically significant improvement over mBERT ($p < 0.05$).

Task	G-XLT						XLT	
Language pair (Q-A)	De-En	De-Es	En-De	En-Es	Es-De	Es-En	De-De	Es-Es
XLM-R	57.8/42.9	41.4/28.5	57.8/45.2	54.4/41.0	49.5/32.2	63.9/43.5	60.8/45.3	64.8/44.3
mBERT	60.1/44.6	53.7/40.0	63.7/50.3	64.1/49.8	57.5/38.7	64.4/43.3	60.7/44.5	63.7/42.2
Ours (mBERT-GSW)	60.9/45.6	55.3*/39.9	65.1/51.2	64.7/51.2	58.9/40.2*	65.1/44.4	61.5/45.5	64.8/43.3
Ours (mBERT, QLM-RR, 10)	62.6*/46.3*	56.4*/39.8	70.6*/56.4*	68.3*/53.2*	62.6*/40.0	67.3*/45.8*	61.9/44.9	63.9/41.8
Ours (mBERT, QLM-RR, 20)	62.3*/45.6	**61.4***/**44.6***	73.2*/58.9*	**71.6***/**57.1***	**65.9***/43.5*	**68.3***/**46.7***	62.3*/46.4*	**66.5***/44.1*
Ours (mBERT-GSW, QLM-RR, 10)	63.5*/46.8*	59.3*/42.7*	72.1*/58.4*	69.2*/53.0*	64.9*/**44.0***	67.7*/46.1*	61.6/44.3	65.2/42.2
Ours (mBERT-GSW, QLM-RR, 20)	**63.7***/**47.3***	60.8*/43.5*	**73.5***/**60.2***	71.5*/57.0*	65.0*/42.7*	68.1*/46.1*	**62.4***/**46.6***	66.1*/**44.8***

is because we strictly pretrain the models with only *bilingual* query-document pairs. In other words, XLT is a zero-shot task with respect to our pretraining data, and it is more difficult to improve on.

We also perform cross-task comparisons (Tables 4 and 5), and observe that when pretrained with the same objectives for the same number of epochs, mBERT-GSW consistently performs better than mBERT on CLIR, but two models perform similarly for QA. Documents ("contexts") in SQuAD and MLQA dataset are paragraphs and are in general shorter than news articles in CLEF. Therefore, pretraining on longer texts does not provide additional gains for QA, as finetuning and testing data have smaller sequence length.

5.3 Ablation Studies and Parameter Analysis

5.3.1 Utilities of pretraining tasks. We conduct experiments to investigate the effect of each pretraining objective on downstream tasks. For efficiency, we use language models that have limit input sequences to first 512 tokens (with self-attention) instead of those that limit inputs to 1024 tokens (with GSW attention). We pretrain two more models, one with only QLM objective, and one with only RR objective. We record all compared models' performance in each available evaluation language pair in both CLIR and CLQA task. We do not observe significant differences in the patterns shown across language pairs, so we select two language pairs per task for demonstration. The results are shown in Figure 4. We use the same evaluation strategies and report performances at different pretraining epochs. Note that the starting points (pretrained epochs = 0) in all subfigures refer to the base mBERT model.

In the case of CLIR, both RR and QLM provide positive gains for retrieval, while the former is much more effective than the latter. This is not surprising considering that RR takes the same form as ad-hoc retrieval using language models. We do observe that there exists mutual complement between these two objectives, as QLM+RR performs significantly better than either alone. For CLQA, again both RR and QLM provide benefits towards this downstream task. However, QLM is more effective compared to RR in this case. The two pretraining objectives are also reciprocal as the LMs pretrained with both objectives perform the best in terms of F1.

In conclusion, we show that QLM and RR have positive influence on downstream cross-lingual retrieval tasks *individually*, and those positive effects are also *additive* such that language models pretrained with both tasks give the best performance.

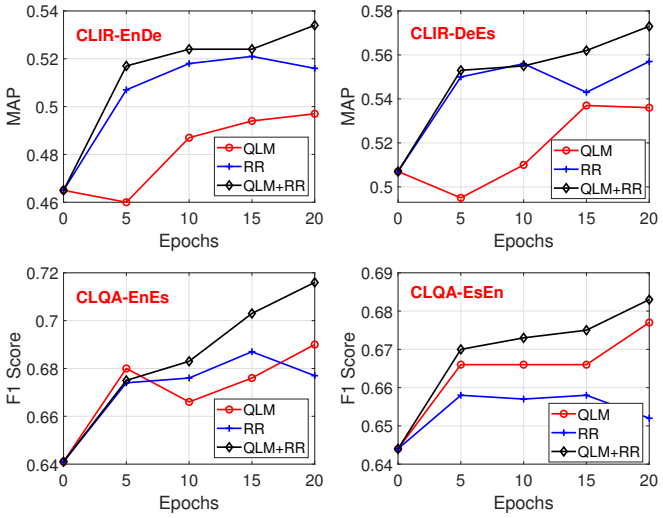

Figure 4: CLIR and CLQA performance of models pretrained with different objectives (QLM, RR, QLM+RR). Epochs=0 refers to base mBERT without RR or QLM pretraining. Best viewed in color.

5.3.2 QLM masking. Two choices distinguish QLM from other pretraining tasks: (i) we only mask query tokens; and (ii) we increase probability of masking from 15% to 30%. Here we conduct controlled experiments to justify our choices. We pretrain five models with only QLM objective under different settings. We regard masking query with 30% probability as a baseline. Two models are trained with different query masking probability (15% and 45%). Other two models are trained with masking probability 30%, but masking different regions (documents and query+document). Their performances on CLEF are reported in Figure 5.

Masking probability: Comparing red, blue and black plots, we observe that $p = 0.15$ is consistently out-performed by higher query masking probability. This indicates that $p = 0.15$ makes the QLM task too easy for the language model to learn cross-lingual query document interactions. Query masking probability $p = 0.45$ performs slights better than $p = 0.30$ when the model is well trained, suggesting that further increasing query masking probability can provide additional gains. **Masking region:** Comparing blue, pink and green plots, it is obvious that masking just document tokens

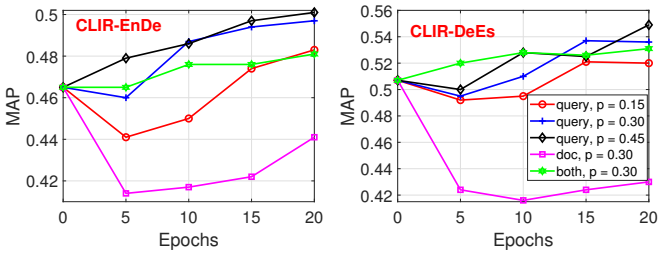

Figure 5: CLIR performance of models pretrained with QLM objective under different parameters (masking region, masking probability). Epochs=0 refers to base mBERT.

greatly hurts CLIR performance at the beginning of pretraining, and cannot recover to the level of vanilla mBERT re-ranker after 20 epochs of pretraining. Masking just query and masking both query and document can spark improvement, but the former performs significantly better when the model is well trained.

In conclusion, we show that masking only query tokens with higher probability (≥0.3) is the better setting for the QLM task. This finding aligns well with our intuitions.

5.3.3 Semantic alignment on sentence and document level. In order to investigate if our proposed cross-lingual retrieval-oriented modeling tasks can promote cross-lingual alignment on more coarse-grained semantics, we evaluate pretrained LMs on two cross-lingual alignment tasks, namely, cross-lingual sentence alignment (**XSA**) and cross-lingual document alignment (**XDA**). XSA/XDA requires that parallel sentences/documents should have embeddings that are proximate in the representation space. We encode a sentence (document) with the LM as "`[cls]` Content `[sep]`", and take the `[cls]` token's last hidden state as the sentence (document) embedding.

Metric. Given a set of sentences $\mathcal{S}$ and a set of corresponding parallel (translation) sentences $\mathcal{T}$, we measure precision at top-1 (P@1), which is defined as

$$P@1 = \frac{1}{|\mathcal{S}|} \sum_{s_i \in \mathcal{S}} \mathbb{1}\left((\underset{j, t_j \in \mathcal{T}}{\operatorname{argmax}} \operatorname{Sim}(s_i, t_j)) = i \right), \tag{2}$$

where $\mathbb{1}$ is the indicator function, and Sim is a function that measures the similarity of two cross-lingual sentences. We use the CSLS measure [28] with neighborhood size set to 10 [5] as the Sim function, which is a modified version of cosine similarity measure. CSLS is widely adopted for evaluating cross-lingual word alignment for two advantages over cosine similarity: (i) CSLS is a symmetric measurement, meaning that switching $\mathcal{S}$ and $\mathcal{T}$ in evaluation does not affect the degree of alignment; and (ii) CSLS can mitigate the hubness problem [15].

Data. Evaluating XSA and XDA requires parallel sentences and documents. For XSA, we adopt the XNLI-15way dataset[11]. It contains 10K sentences that are manually translated into 15 languages. We select the four languages that overlap with our pretraining languages and form six evaluation pairs. For XDA, we use the United Nations Parallel Corpus[12], which contains 86,307 documents in six languages that are official United Nations languages. Out of the four languages in our pretraining languages, German is not in the corpus. Therefore we use English, French and Spanish data and form three language pairs. Same as in Section 5.3.1 and 5.3.2, we use LMs with input limit to 512 tokens for efficiency.

Results. We report the XSA and XDA performance of base mBERT, as well as cross-lingual LMs that have been additionally pretrained with our retrieval-oriented modeling tasks for 20 epochs in Table 6 and Table 7. We observe that base mBERT generates poor cross-lingual sentence and document alignment, except between Spanish and French documents. In terms of XSA, QLM and RR can bring improvement upon base mBERT. The improvement is more prominent with QLM, and that the effect is not additive on XSA. In terms of XDA, QLM and RR work similarly well, and combining two modeling tasks together can spark further improvement on cross-lingual document alignment. The differences between XSA and XDA is understandable considering that QLM is more focused on sentence-level semantics (sentence completion given foreign document) and that RR is more focused on document-level (ranking documents with respect to foreign sentence). In all circumstances, pretraining the model with either RR or QLM on the Wiki weak-supervision data can significantly improvement cross-lingual coarse-grained semantics alignment. This might lead to improvement on more applications beyond cross-lingual retrieval.

Table 6: Cross-lingual sentence alignment (XSA) results on XNLI-15way dataset. Numbers are P@1 in percentage (%).

Language pairs	De&En	De&Es	De&Fr	En&Es	En&Fr	Es&Fr
Base mBERT	1.9	4.3	6.3	1.6	2.2	7.4
+ QLM	**27.4**	**51.9**	**52.3**	**53.2**	**50.9**	**63.5**
+ RR	16.6	25.8	28.4	33.8	22.5	46.0
+ QLM&RR	22.7	36.4	45.4	45.4	43.1	57.9

Table 7: Cross-lingual document alignment (XDA) results on MultiUN dataset. Numbers are P@1 in percentage (%).

Language pairs	En&Es	En&Fr	Es&Fr
Base mBERT	2.6	5.2	30.3
+ QLM	24.9	23.3	61.0
+ RR	19.1	22.5	65.1
+ QLM&RR	**40.3**	**43.1**	**81.5**

5.3.4 Effect of window size in GSW attention. An assumed key parameter in the global+sliding window (GSW) attention is the window size w. In the context of cross-lingual retrieval, it represents the number of neighboring tokens a document token can "attend" to in a single Transformer layer. Although GSW is theoretically superior to full self-attention in terms of efficiency (linear versus square), a large window size might void the effort in practice. Therefore, it is important to evaluate how the setting of window size might influence the performance on desired tasks. To that end, we choose four different window size $w = \{32, 64, 128, 256\}$ for experiments. For all four models, we set input sequence limitation to 1024 tokens, pretrain them on English-French part of the weak-supervision data (described in Section 4.1) on QLM+RR pretraining modeling tasks,

[11] https://github.com/facebookresearch/XNLI
[12] https://conferences.unite.un.org/uncorpus

Table 8: The effect of window size w in GSW attention on CLIR performance. Numbers are MAP.

Languages (Q→D)	En→Fr	Fr→En
$w = 32$	0.512	0.509
$w = 64$	0.510	0.506
$w = 128$	0.516	0.512
$w = 256$	0.509	0.503

and perform CLIR evaluation on CLEF dataset (as in Section 5.1) in two directions (En→Fr and Fr→En). The results are reported in Table 8. We observe no statistically significant differences on downstream CLIR performance caused by different window size w in GSW attention. Taking into account the randomness of model training, we conclude that window size w in the GSW attention has no obvious impact on the CLIR task. In practice, one is suggested to prioritize longer sequence length over larger window size when facing a trade-off.

6 CONCLUSION

In this work, we show that the absence of cross-lingual passage-level relevance data and the lack of proper query-document style pretraining are key reasons for the inferior performance in adopting multi-lingual language models for CLIR. To overcome such difficulties, we introduce two novel pretraining objectives to improve Transformer based cross-lingual language models for retrieval tasks. We also propose building fine-grained cross-lingual query-document style weak-supervision data from multilingual Wiki to support large-scale pretraining. We employ global+sliding-window attention to allow language models to encode much longer documents in all three stages (pretraining, finetuning, and inference) efficiently. Extensive experiments demonstrate the effectiveness of our contributions on both cross-lingual ad-hoc retrieval and cross-lingual extractive question answering. Detailed ablation studies justify our modeling choices and parameter selections. We also discover that our model can significantly improve coarse-grained semantic alignment across languages, which might lead to a wider range of applications beyond retrieval.

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada, 451–462.

[2] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[3] Rohit G. Bharadwaj and Vasudeva Varma. 2011. Language independent identification of parallel sentences using Wikipedia. In *Proceedings of the 20th International Conference on World Wide Web (WWW) (Companion Volume)*. Hyderabad, India, 11–12.

[4] Hamed R. 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 (SIGIR)*. Virtual Event, China, 9–18.

[5] Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual Alignment of Contextual Word Representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia.

[6] Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training Tasks for Embedding-based Large-scale Retrieval. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia.

[7] Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-Powered Representation Learning for Cross-Lingual Sentiment Classification. In *Proceedings of the World Wide Web Conference (WWW)*. San Francisco, CA, 251–262.

[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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Online, 8440–8451.

[9] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 7057–7067.

[10] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging Cross-lingual Structure in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Online, 6022–6034.

[11] 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 Information Retrieval (SIGIR)*. Toronto, Canada, 338–344.

[12] Gerard de Melo. 2017. Inducing Conceptual Embedding Spaces from Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web (WWW) (Companion Volume)*. Perth, Australia, 43–50.

[13] 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 (NAACL-HLT)*. Minneapolis, MN, 4171–4186.

[14] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC Complex Answer Retrieval Overview. In *Proceedings of The Twenty-Sixth Text REtrieval Conference (TREC)*. Gaithersburg, MD.

[15] Georgiana Dinu and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. San Diego, CA.

[16] Hongliang Fei and Ping Li. 2020. Cross-Lingual Unsupervised Sentiment Classification with Multi-View Transfer Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Online, 5759–5771.

[17] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM)*. Indianapolis, IN, 55–64.

[18] Jiafeng Guo, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2019. MatchZoo: A Learning, Practicing, and Developing System for Neural Text Matching. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Paris, France, 1297–1300.

[19] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Online, 8342–8360.

[20] Djoerd Hiemstra and Franciska de Jong. 1999. Disambiguation Strategies for Cross-Language Information Retrieval. In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*. Paris, France, 274–293.

[21] Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. 2020. Local Self-Attention over Long Text for Efficient Document Retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*. Virtual Event, China, 2021–2024.

[22] Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. 2020. Interpretable & Time-Budget-Constrained Contextualization for Re-Ranking. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*. Santiago de Compostela, Spain, 513–520.

[23] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 2485–2494.

[24] Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos G. Karakos, and Lingjun Zhao. 2020. Cross-lingual Information Retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS@LREC)*. Marseille, France, 26–31.

[25] Martin Josifoski, Ivan S. Paskov, Hristo S. Paskov, Martin Jaggi, and Robert West. 2019. Crosslingual Document Embedding as Reduced-Rank Ridge Regression. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM)*. Melbourne, Australia, 744–752.

[26] Armand Joulin, Piotr Bojanowski, Tomás Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, 2979–2984.

[27] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. San Diego, CA.

[28] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. Vancouver, Canada.

[29] Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Online, 7315–7330.

[30] Robert Litschko, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2018. Unsupervised Cross-Lingual Information Retrieval Using Monolingual Data Only. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*. Ann Arbor, MI, 1253–1256.

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[32] 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 IR Research (ECIR)*. Lisbon, Portugal, 246–254.

[33] 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 (SIGIR)*. Paris, France, 1101–1104.

[34] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. Online.

[35] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *Proceedings of the 18th International Conference on World Wide Web (WWW)*. Madrid, Spain, 1155–1156.

[36] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[37] Tiziano Piccardi, Michele Catasta, Leila Zia, and Robert West. 2018. Structuring Wikipedia Articles with Section Recommendations. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*. Ann Arbor, MI, 665–674.

[38] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Melbourne, Australia, 275–281.

[39] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[40] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX, 2383–2392.

[41] Andreas Rücklé, Krishnkant Swarnkar, and Iryna Gurevych. 2019. Improved Cross-Lingual Question Retrieval for Community Question Answering. In *Proceedings of The World Wide Web Conference (WWW)*. San Francisco, CA, 3179–3186.

[42] Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A Survey of Cross-lingual Word Embedding Models. *J. Artif. Intell. Res.* 65 (2019), 569–631.

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

[44] Daniel Stein, Dimitar Sht. Shterionov, and Andy Way. 2019. Towards language-agnostic alignment of product titles and descriptions: a neural approach. In *Proceedings of the World Wide Web Conference (WWW), (Companion Volume)*. San Francisco, CA, 387–392.

[45] Ferhan Türe and Jimmy J. Lin. 2013. Flat vs. hierarchical phrase-based translation models for cross-language information retrieval. In *Proceedings of the 36th International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*. Dublin, Ireland, 813–816.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS)*. Long Beach, CA, 5998–6008.

[47] Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Santiago, Chile, 363–372.

[48] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Shinjuku, Tokyo, 55–64.

[49] Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Simple applications of BERT for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972* (2019).

[50] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 5754–5764.

[51] Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Applying BERT to Document Retrieval with Birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 19–24.

[52] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 3488–3494.

[53] Puxuan Yu and James Allan. 2020. 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)*. Virtual Event, China, 1637–1640.