A General-Purpose Multilingual Document Encoder

Onur Galoğlu[1] **Robert Litschko**[2*] **Goran Glavaš**[3]

[1]Independent Researcher
[2]MaiNLP, Center for Information and Language Processing (CIS), LMU Munich, Germany
[3]CAIDAS, University of Würzburg

Abstract

Massively multilingual pretrained transformers (MMTs) have tremendously pushed the state of the art on multilingual NLP and cross-lingual transfer of NLP models in particular. While a large body of work leveraged MMTs to mine parallel data and induce bilingual document embeddings, much less effort has been devoted to training general-purpose (massively) multilingual document encoder that can be used for both supervised and unsupervised document-level tasks. In this work, we pretrain a massively multilingual document encoder as a hierarchical transformer model (HMDE) in which a shallow document transformer contextualizes sentence representations produced by a state-of-the-art pretrained multilingual sentence encoder. We leverage Wikipedia as a readily available source of comparable documents for creating training data, and train HMDE by means of a cross-lingual contrastive objective, further exploiting the category hierarchy of Wikipedia for creation of difficult negatives. We evaluate the effectiveness of HMDE in two arguably most common and prominent cross-lingual document-level tasks: (1) cross-lingual transfer for topical document classification and (2) cross-lingual document retrieval. HMDE is significantly more effective than (i) aggregations of segment-based representations and (ii) multilingual Longformer. Crucially, owing to its massively multilingual lower transformer, HMDE successfully generalizes to languages unseen in document-level pretraining. We publicly release our code and models.[1].

1 Introduction

Massively multilingual Transformers (MMTs) such as XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021) have drastically pushed the state-of-the-art in multilingual NLP, especially for medium-resourced languages included in their pretraining, enabling effective cross-lingual transfer of task-specific NLP models from languages with plenty of training data to languages with little or no annotated task data. Being standard transformer-based language models, MMTs process text linearly – as a flat sequence of tokens, which has – in monolingual contexts – been shown suboptimal for document-level tasks (e.g., document classification or retrieval) for two main reasons: (1) it does not correspond to the hierarchical nature of document organization – documents are sequences of (presumably meaningfully ordered) paragraphs, which are in turn sequences of sentences (Zhang et al., 2019; Glavaš and Somasundaran, 2020), and (2) representing documents longer than the MMTs maximal input length requires either document trimming, which leads to loss of potentially task-relevant information, or segmentation, which leading to context fragmentation (Ding et al., 2021).

A number of models that produce document-level representations have been proposed, albeit predominantly in the monolingual (English) realm, with two prominent lines of work. **(1)** Hierarchical encoders (Pappas and Popescu-Belis, 2017; Pappagari et al., 2019; Zhang et al., 2019; Yang et al., 2020; Glavaš and Somasundaran, 2020; Chalkidis et al., 2022) typically contextualize sentence-level representations with additional document-level parameters (e.g., an additional, document-level transformer). These document-level parameters of the encoder, added on top of a pretrained language model like BERT (Devlin et al., 2019), are typically trained on large task-specific datasets, ranging from document classification (Pappagari et al., 2019) to summarization (Zhang et al., 2019) and segmentation (Glavaš and Somasundaran, 2020). Task-specific training of document-level parameters impedes the transfer of such encoders to other tasks. **(2)** Sparse attention models (Child et al., 2019; Zaheer et al., 2020; Beltagy et al., 2020; Tay et al., 2020) modify the attention mechanism in

[*] Work done while at University of Mannheim
[1]https://github.com/ogaloglu/pre-training-multilingual-document-encoders

Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL), pages 37–49
December 7, 2023 ©2023 Association for Computational Linguistics

order to reduce its computational complexity and consequently be able to encode longer texts. Although flat long-text encoders do not model the hierarchical nature of documents, they allow for flat encoding of substantially longer documents.

In this work, we demonstrate the benefits of hierarchical document representations in multilingual context. We propose to train a hierarchical transformer model (HMDE), coupling (i) a pretrained multilingual sentence encoder as a lower encoder with (ii) an upper transformer that contextualizes sentence representations against each other and from which we derive document representations. Unlike in monolingual setup, where task-specific data is commonly used to train the parameters of the upper transformer (Zhang et al., 2019; Glavaš and Somasundaran, 2020), we exploit the fact that in the multilingual context one can leverage cross-lingual document alignments to guide the *pretraining* of the document encoder, i.e., its upper transformer. To this end, we leverage Wikipedia as a readily available source of quasi-parallel documents, and additionally exploit its hierarchy of categories to create hard negative examples for our contrastive pretraining objective.

We evaluate HMDE in two arguably most prominent (cross-lingual) document-level tasks: (1) cross-lingual transfer for document classification (XLDC) and (2) cross-lingual document retrieval (CLIR). For XLDC, as a supervised task, we fine-tune HMDE on English task-specific data; in CLIR, in contrast, we leverage HDME in an unsupervised fashion, using it to produce static document embeddings (and its lower transformer to produce query embeddings). HDME exhibits performance superior to that of competitive models – MMTs with sliding window and multilingual Longformer (Yu et al., 2021; Sagen, 2021). Crucially, HMDE generalizes well to languages unseen in its document-level pretraining. Our further analyses offer additional insights: (i) that it is important to allow updates from document-level training to propagate to the sentence-level encoder (i.e., not to freeze the parameters of the pretrained sentence encoder) and (ii) that the size of the document-level pretraining corpora matters more than its linguistic diversity (i.e., number of languages it encompasses).

2 Hierarchical Multilingual Encoder

The HMDE architecture, illustrated in Figure 1, is similar to that of hierarchical document encoders trained monolingually in task-specific training (e.g., (Glavaš and Somasundaran, 2020)): a sentence-level (lower) encoder produces sentence embeddings from tokens, whereas the document-level (upper) transformer yields document representation from a sequence of sentence embeddings. We initialize the lower transformer with the pretrained weights of a multilingual sentence encoder (Feng et al., 2022), and train the whole model via a bi-encoder configuration (also known as Siamese architecture) – where we compute a similarity score between representations of two documents produced independently with HDME – using a cross-lingual contrastive objective with both in-batch and hard negatives (Oord et al., 2018).

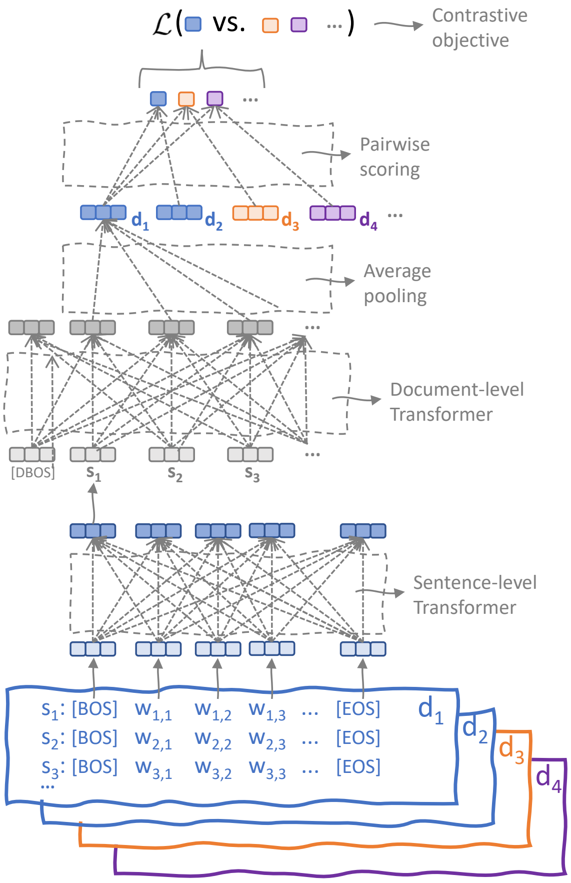
$\mathcal{L}$(vs. ...)
Contrastive objective
Pairwise scoring
d1 d2 d3 d4 ...
Average pooling
Document-level Transformer
[DBOS] s1 s2 s3
Sentence-level Transformer
s1: [BOS] w1,1 w1,2 w1,3 ... [EOS]
s2: [BOS] w2,1 w2,2 w2,3 ... [EOS]
s3: [BOS] w3,1 w3,2 w3,3 ... [EOS]
d1 d2 d3 d4

Figure 1: Illustration of HDME: hierarchical transformer architecture coupled with a cross-lingual contrastive objective. Document colors indicate the Wikipedia concepts: d_1 and d_2 are the pages of the same concept (e.g., New York) in two different languages, L_1 and L_2; documents d_3 and d_4 are pages of other concepts in L_1. The pair (d_1, d_2) is a positive pair (i.e., same concept) for the contrastive training objective and pairs (d_1, d_3) and (d_1, d_4) are corresponding negative pairs (i.e., different concepts).

2.1 Hierarchical Encoding

The role of the sentence-level (lower) transformer is to produce sentence representations from sequences of tokens. Because of this, we initialize it with the pretrained weights (including subword embeddings) of LaBSE (Feng et al., 2022), a state-of-the-art multilingual sentence encoder.[2] The sentence embedding is the transformed representation of the special beginning-of-sequence (BOS) token. The sequence of sentence embeddings obtained with the sentence-level transformer is then forwarded to the document-level (upper) transformer, which mutually contextualizes them, prepended with a special document-level beginning-of-sequence token (DBOS, with a randomly initialized embedding). We derive the document representation by average-pooling contextualized sentence embeddings (i.e., output of the last layer of the document-level transformer).[3]

2.2 Multi- and Cross-Lingual Objective

Our training dataset consists of Wikipedia pages written in one of n languages (see §3.1 for details on the creation of different training datasets): let $L = L_1, L_2, \ldots, L_n$ denote our set of training languages. In each training step, we select a batch of N documents pairs, $\{(d_1^{(1)}, d_2^{(1)}), \ldots, (d_1^{(N)}, d_2^{(N)})\}$, where $d_1^{(i)}$ and $d_2^{(i)}$ are Wikipedia pages of the same concept but in two different languages, L_k and $L_m \in L$. Each of the documents $d_1^{(i)}$ (i.e., first document of each pair) is additionally paired with a document $d_{neg}^{(i)}$ – a document in the same language L_k as $d_1^{(i)}$ and from the same Wikipedia category – representing a *hard negative* for $d_1^{(i)}$ (see §3.1 for details). We then compute and minimize a variant of the popular InfoNCE loss (Oord et al., 2018) that incorporates hard negatives, treating all other batch documents $d_2^{(j)}$ as in-batch (easy) negatives for $d_1^{(i)}$:

$$\mathcal{L} = -\sum_{i=1}^{N} \Bigg[\frac{1}{\tau} s(\mathbf{d}_1^{(i)}, \mathbf{d}_2^{(i)}) - \log\left(e^{s(\mathbf{d}_1^{(i)}, \mathbf{d}_{neg}^{(i)})/\tau} + \sum_{j=1}^{N} e^{s(\mathbf{d}_1^{(i)}, \mathbf{d}_2^{(j)})/\tau} \right) \Bigg] \quad (1)$$

with $\mathbf{d} \in \mathbb{R}^h$ as the embedding of d, i.e., the output of the document-level transformer (and h as the hidden size of upper transformer), $s(\mathbf{d}_i, \mathbf{d}_j)$ as the scoring function capturing similarity between the two document embeddings, and τ as the hyperparameter (the so-called temperature) of the InfoNCE loss. Following common practice, we use cosine similarity as the scoring function s.

Note that the loss we compute is both multilingual and cross-lingual: documents $d_1^{(i)}$ come from any of the $|L|$ languages, and positive pairs $(d_1^{(i)}, d_2^{(i)})$ are cross-lingual. Among the in-batch negatives, there will be cross-lingual as well as monolingual pairs (when $d_1^{(i)}$ and $d_2^{(j)}$ happen to be documents written in the same language). Our hard negatives are, by design, always monolingual pairs. While one could create cross-lingual hard negatives in the same manner (e.g., by pairing the English article *"France"* with an Italian article *"Svizzera"* (Switzerland) that covers another concept from the same category *"Country"*), monolingual hard negatives should be *harder* because the two document representations will originate from the same language-specific subspace of the embedding space of the lower (multilingual) transformer (Cao et al., 2020; Wu and Dredze, 2020).

3 Experimental Setup

We first describe how we created the multilingual dataset for HMDE pretraining from Wikipedia (§3.1). We then briefly describe the two evaluation tasks – cross-lingual transfer for document classification and cross-lingual information retrieval – and their respective datasets (§3.2), following with the description of the baselines – a multilingual sentence encoder with a sliding window and a multilingual Longformer (Yu et al., 2021; Sagen, 2021) (§3.3). We provide training and optimization details for all models in the Appendix A.1.

3.1 Data Creation

Wikipedia has been leveraged as a suitable source for mining comparable and parallel corpora for decades (Ni et al., 2009; Plamadă and Volk, 2013; Schwenk et al., 2021, *inter alia*). We add to the body of work that exploits Wikipedia as a massively multilingual text resource by using it to build pretraining data for HMDE. Concretely, for a set of languages $L = \{L_1, L_2, \ldots, L_n\}$, we first fetch

[2] We load LaBSE weights from HuggingFace: `https://huggingface.co/sentence-transformers/LaBSE`

[3] We preliminarily also experimented with the contextualized vector of the DBOS token as the document representation, but that consistently led to lower performance.

monolingual portions from the Wiki-40B corpus.[4] We then identify articles in different languages that are about the same concept (via the `wikidata_id` field) and keep only those concepts for which pages are found in at least two languages from L. For each such concept with pages $p_1, p_2, \ldots, p_m$ in m different languages, we create all possible cross-lingual pairs of articles (p_i, p_j) covering the same concept. For each pair (p_i, p_j), we then leverage Wikipedia metadata – namely mapping of Wikipedia pages into its hierarchy of categories – to select an article n_i from the same monolingual Wikipedia as p_i (i.e., written in the same language as p_i) that belongs to (at least one) same Wikipedia category as p_i. This yields triples (p_i, p_j, n_i) from which we create cross-lingual positives (p_i, p_j) and their corresponding monolingual hard negatives (p_i, n_i) for our contrastive objective (see §2.2).

On the one hand, the quality of MMTs' representations of a particular language depends on the size of the pretraining corpora of that language (Hu et al., 2020; Lauscher et al., 2020). On the other hand, multilingual model training with instances from linguistically diverse languages may generalize better to unseen languages (Chen et al., 2019; Ansell et al., 2021). Most resourced languages, however, tend to be Indo-European (Joshi et al., 2020), putting corpus size and linguistic diversity at odds. We thus create two different datasets, each emphasis one of these two aspects: (1) XLW-4L is built starting from four high-resource Indo-European languages: English, German, French, and Italian; (12) XLW-12L is built starting from a set of 12 linguistically diverse languages: English, French, Russian, Japanese, Chinese, Hungarian, Finnish, Arabic, Persian, Turkish, Greek, and Malay. With 1.1M triples (p_i, p_j, n_i), XLW-4L is almost twice as large as XLW-12L (which encompasses 592K triples), despite encompassing three times fewer languages: this is primarily because there are many more shared concepts between large Wikipedias of XLW-4L (e.g., German and Italian) than between smaller Wikipedias of XLW-12L (e.g., Turkish and Malay).[5]

3.2 Evaluation Tasks and Datasets

HMDE is meant to be a general-purpose multilingual document encoder. It thus needs to be useful both (1) when fine-tuned for a supervised document-level task, and (2) as a standalone document encoder. We thus evaluate HMDE in (1) zero-shot cross-lingual transfer for supervised document classification (XLDC) and (2) unsupervised cross-lingual document retrieval (CLIR).

XLDOC. Regular MMTs (e.g., mBERT or XLM-R) are primarily used in zero-shot cross-lingual transfer for supervised NLP tasks: an MMT fine-tuned on task-specific training data in a resource-rich language is used to make predictions for language(s) without task data. We evaluate HMDE in exactly the same zero-shot cross-lingual transfer setup, only for a document-level task – topical document classification. We fine-tune HMDE in the standard manner, by stacking a softmax classifier on top the output of the document-level encoder. With $\mathbf{d}$ as HDME's encoding of the input document d, classifier's prediction is computed as:

$$\mathbf{y} = \mathit{softmax}\left(\mathbf{W} \cdot \mathbf{d} + \mathbf{b}\right) \qquad (2)$$

with $\mathbf{W} \in \mathbb{R}^{C \times h}$ and $\mathbf{b} \in \mathbb{R}^{C}$ as classifier's trainable parameters (and C as the number of classes).

We fine-tune HMDE on the English training portion of the MLDOC dataset (Schwenk and Li, 2018) and evaluate its performance on the test portions of all other (target) languages. MLDOC is a subset of the Reuters Corpus Volume 2 (RCV2), with training, development, and test portions in 8 languages (English, Spanish, German, French, Italian, Russian, Japanese and Chinese), consisting of 1000, 1000, and 4000 documents, respectively. News stories are categorized into $C = 4$ semantically closely related classes (*Corporate/Industrial*, *Economics*, *Government/Social*, and *Markets*).

CLIR. We evaluate the effectiveness of HMDE as a standalone document encoder in an unsupervised cross-lingual document retrieval task: queries (short text) in one language are fired against a collection of documents written in another language. We adopt a simple retrieval model: we rank documents in decreasing order of cosine similarity of their embeddings $\mathbf{d}$, produced by the HMDE, with the embedding $\mathbf{q}$ of the query, $\cos(\mathbf{d}, \mathbf{q})$. We obtain the query embedding $\mathbf{q}$ by encoding the query only with HMDE's lower (sentence-level) transformer: $\mathbf{q}$ is the transformed representation of the beginning-of-sequence ([BOS]) token.

We carry out the evaluation on CLEF-2003,[6] a popular CLIR benchmark, including the following

[4] Available in Tensorflow datasets: `https://www.tensorflow.org/datasets/catalog/wikipedia`

[5] Per-language statistics of the datasets are in the Appendix.

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

languages: English (EN), German (DE), Italian (IT), Finnish (FI) and Russian (RU). Following prior work (Glavaš et al., 2019; Litschko et al., 2022), we evaluate HMDE on 9 language pairs (with first language being the query language): EN-FI, DE, IT, RU, DE-FI, IT, RU, FI-IT, RU. For each language pair we work with 60 queries and document collections of following sizes: RU – 17K, FI – 55K, IT – 158K, and DE – 295K.

3.3 Baseline Models

There are two main alternatives to hierarhical (long) document encoding. The first is to (i) fragment the document into smaller segments, (ii) encode each segment with a regular pretrained MMT (e.g., vanilla MMT like XLM-R or a multilingual sentence encoder like LaBSE), and (iii) aggregate the document representation from the embeddings of segments. The second is to train a multilingual sparse-attention encoder, akin to (Sagen, 2021).

MMT with a Sliding Window (LaBSE-Seg). For fair comparison, we use LaBSE (Feng et al., 2022) – the same pretrained MMT that we use for the initialization of the lower transformer in HMDE – to independently encode overlapping segments of the input document. We break down the document into segments of length N_S tokens. Following Dai et al. (2022), who find that overlapping segments alleviate the context fragmentation problem, we make adjacent segments overlap in $N_S/3$ tokens. After encoding each segment with LaBSE, we average-pool the document representation $\mathbf{d}$ from the set of segment embeddings. In XLDX (topical document classification) this average of segment embeddings is fed into the classification head. In CLIR, it is compared with the LaBSE encoding of the query.

Multilingual Longformer (mLongformer). Longformer architecture (Beltagy et al., 2020) combines local-window attention with global attention, resulting in a hybrid attention mechanism, the memory requirements of which scale linearly with the input length. Beltagy et al. (2020) additionally propose multi-step procedure for initializing Longformer's parameters based on the parameters of a pretrained regular transformer (e.g., in the case of monolingual English Longformer from RoBERTa (Liu et al., 2019)) and then further train the Longformer via masked language modeling (MLM). We train the multilingual Longformer following the same procedure: for fair comparison with HMDE, we initialize its parameters from the parameters of LaBSE and carry out the additional MLM training on XLW-4L, the same corpus on which we train HMDE.

4 Results and Discussion

We first report and discuss the main results we obtain with HMDE on XLDC and CLIR (in §4.1). In a series of follow-up experiments, we further analyze key design choices for HMDE (§4.2).

4.1 Main Results

Cross-lingual Document Classification. Table 1 compares HMDE trained on XLW-4L against several standard and long document multilingual encoders: besides the baselines introduced in §3.3, for completeness we add the results for vanilla LaBSE (i.e., without sliding over the long document) and models based on XLM-R and mBERT reported by Dong et al. (2020) and Zhao et al. (2021), respectively. Expectedly, all long-document encoders outperform all of the standard MMTs. mLongformer and HMDE generally exhibit similar performance, surpassing the performance of segmentation-based LaBSE-Seg for virtually all languages. Comparable performance of mLongformer and HMDE suggests that in the presence of task-specific finetuning data it does not really matter whether we aggregate document representations in a flat or hieratrchical fashion. What is particularly encouraging is that both HDME and mLongformer exhibit strong performance for languages that they did not observe in document-level pretraining: Spanish, Russian, Japanese, and Chinese.[7,8]

Cross-lingual Retrieval. The results for unsupervised CLIR are shown in Table 2. Like in XLDC, we additionally report the results for LaBSE that encodes only the beginning of the document (without sliding) as well as for mBERT, reported by Litschko et al. (2022). CLIR, in which multilingual transformers are used as standalone document encoders without any task-specific fine-tuning, tell a very different story from supervised XLDC results. HMDE drastically outperforms mLongformer, indicating that, much like the vanilla MMTs, mLongformer requires fine-tuning and cannot encode reli-

[7] LaBSE, with whose parameters both HMDE and mLongofrmer were initialized before document-level pretraining, however, was exposed to all of these languages in its own sentence-level pretraining.

[8] Performance across languages *not* directly comparable as MLDOC test sets are not parallel across languages.

Model	En	Es	De	Fr	It	Ru	Ja	Zh	AVG
Standard Multilingual Transformers									
LaBSE	95.5	79.0	89.6	87.2	76.8	63.9	**80.8**	86.1	82.4
XLM-R (Dong et al., 2020)	93.0	84.6	**92.5**	87.1	73.2	68.9	78.2	85.8	83.0
mBERT (Zhao et al., 2021)	**96.9**	81.9	88.3	83.1	74.1	72.3	74.6	84.4	82.0
Multilingual Long Document Encoders									
LaBSE-Seg	94.0	82.9	90.2	89.9	78.1	71.9	75.5	88.4	84.0
mLongformer (XLW-4L)	95.8	**87.0**	93.4	91.9	**80.6**	71.7	79.5	88.5	86.1
HMDE (XLW-4L)	95.4	85.6	91.2	**92.0**	78.5	**83.9**	76.3	**89.5**	**86.8**

Table 1: Performance of HDME compared against standard MMTs and baseline multilingual long-document encoders on supervised topical document classification (MLDOC). Performance (except En) for zero-shot cross-lingual transfer: all models are fine-tuned only on English training data. **Bold**: best performance in each column.

Model	En–Fi	En–It	En–Ru	En–De	De–Fi	De–It	De–Ru	Fi–It	Fi–Ru	AVG
Standard Multilingual Transformers										
LaBSE	.247	.224	.131	.138	.247	.214	.135	.211	.125	.186
mBERT (Litschko et al., 2022)	.145	.146	**.167**	.107	.151	.116	**.149**	.117	.128	.136
Multilingual Long Document Encoders										
LaBSE-Seg	.243	.169	.107	.194	.268	.178	.104	.153	.014	.159
mLongformer (XLW-4L)	.150	.088	.094	.082	.190	.072	.120	.097	.091	.109
HMDE (XLW-4L)	**.380**	**.282**	.141	**.326**	**.352**	**.259**	.130	**.238**	**.129**	**.249**

Table 2: Performance of HDME compared against standard MMTs and baseline multilingual long-document encoders on unsupervised cross-lingual document retrieval (CLEF-2003). **Bold**: best performance in each column.

ably encode documents “out of the box”. HMDE also substantially outperforms LaBSE-Seg, the long-document encoder based on sliding LaBSE over the document. Interestingly, vanilla LaBSE, which encodes only the beginning of the document, also outperforms its sliding counterpart LaBSE-Seg, which is exposed to the entire document. We believe that this is because (1) in CLEF, retrieval-relevant information often occurs at the beginnings of documents and in such cases (2) LaBSE-Seg’s average-pooling over all document segments then dilutes the encoding of query-relevant content. Importantly, HMDE in CLIR also seems to generalize very well to languages unseen in its document-level pretraining (in particular for Finnish documents).

4.2 Further Analysis

We next empirically examine how different choices in HDME’s design and pretraining affect its performance, focusing on: (i) linguistic diversity and size of the pretraining corpus (XLW-4L vs. XLW-12L), (ii) freezing of the lower transformer (i.e., LaBSE weights) after initialization, and (iii) initializing it with the weights of XLM-R as the standard MMT (vs. initialization with LaBSE as the sentence encoder). We provide a further ablations on document segmentation (sentences vs. token sequences ignorant of sentence boundaries) in the Appendix A.2.

Pretraining Data: Linguistic Diversity vs. Size. As discussed in §3.1, we prepare two different corpora for HMDE pretraining: XLW-4L, which is larger (1.1M instances) but encompasses only four major Indo-European languages and XLW-12L, which is smaller (590K instances) but has documents from a set of 12 linguistically diverse languages. To control for the size, and assess the effect of linguistic diversity alone, we randomly downsample XLW-4L, creating a 4-language dataset XLW-4L-S that matches in size XLW-12L. Figure 2 shows the downstream performance of HMDE when pretrained on each of these three datasets.

Comparison between XLW-4L and XLW-4L-S (same languages, different dataset size) shows that

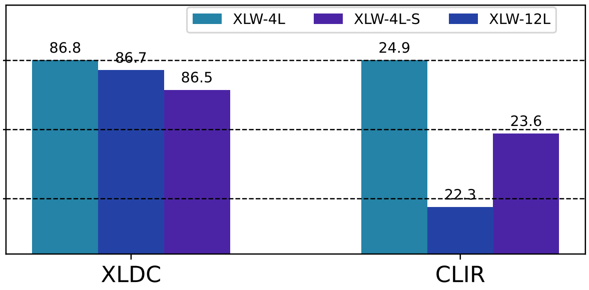

Figure 2: Performance of HMDE when pretrained on different datasets. Results are averages across all test languages (XLDC) and language pairs (CLIR).

our flavor of cross-lingual contrastive pretraining (§2.2) leads to a fairly sample-efficient pre-training: cutting the training data almost in half leads to small performance drops (mere 0.3 accuracy points in XLDC; 1.3 MAP points in CLIR). Comparison between XLW-4L-S and XLW-12L (same size, different language sets) quantifies the role of linguistic diversity in pretraining. Somewhat surprisingly, the more linguistically diverse pretraining on XLW-12L does not bring better performance compared to "Indo-European-only" pretraining on XLW-4L-S: while they perform comparably on XLDC, more diverse pretraining (XLW-12L) leads to worse CLIR performance (-1.3 MAP points on average). We hypothesize that this is due to higher-quality of representation of the four Indo-European languages (EN, DE, FR, IT) in LaBSE (owing to their over-representation in LaBSE's pretraining), with which we initialize the lower transformer of HMDE. We find this result to be particularly encouraging, as – together with the observation that HMDE generalizes well to languages unseen in its document-level pretraining – it suggests that document-level pretraining itself does not necessarily need to be massively multilingual in order to yield successful massively multilingual document encoders.

Lower Transformer. We next investigate two aspects of the lower-transformer: (1) with which weights to initialize it and (2) whether it pays off to update its parameters during the document-level pretraining. For the former, we compare our default LaBSE-based initialization (with LaBSE as a sentence-specialized multilingual encoder) against the initialization with weights of XLM-R, as the vanilla multilingual MMT. To answer the latter, we additionally train HMDE by freezing its lower transformer in document-level pretraining. Table 3 summarizes the results of these ablations.

Model	Updates	XLDC	CLIR
HMDE-LaBSE	*Updated*	**86.8**	**0.249**
HMDE-LaBSE	*Frozen*	85.9	0.167
HMDE-XLM-R	*Updated*	83.9	0.135

Table 3: HMDE results for different choices w.r.t. to initialization and training of the lower transformer. Training for all three variants carried out on XLW-4L. Results are averages across all test languages (XLDC) and language pairs (CLIR).

While freezing the lower transformer after initialization leads to much faster training, it results in poorer document encoder, especially if used for standalone document encoding, without task-specific fine-tuning[9] (HMDE-LaBSE *Updated* vs. *Frozen*; 1 accuracy point drop in XLDC vs. 8 MAP points drop in CLIR). Initializing HDME's lower transformer with LaBSE weights leads to much better downstream performance compared to initialization with XLM-R which is not specialized for sentence-level semantics.

5 Related Work

We position our contributions w.r.t. three related lines of work: (1) pretraining long-document encoders, (2) self-supervised pretraining for retrieval, and (3) mining parallel documents.

Long-Document Encoders. Hierarchical (Zhang et al., 2019; Yang et al., 2020; Glavaš and Somasundaran, 2020) and sparse-attention-based encoders (Beltagy et al., 2020; Zaheer et al., 2020; Tay et al., 2020) already discussed in §1 account for the vast majority of long-document encoding approaches. Dai et al. (2022) extensively compare Longformer (Beltagy et al., 2020) against hierarchical transformers on various long-document classification tasks, showing that the latter exhibit slightly better performance, especially if the lower encoder encodes overlapping segments. Ding et al. (2021) propose a different, segmentation-based model based on recurrence transformers (Dai et al., 2019), designed to remedy for context fragmentation with a retrospective feed mechanism: each segment is encoded twice – after initial left-to-right segment with a recurrent transformer, segment representations are further mutually contextualized

[9] The parameters of the lower-transformer are always updated in XLDC fine-tuning, even if we froze them in document-level pretraining.

bidirectionally. Their training couples MLM-ing with a segment reordering objective.

The vast majority of work on pretraining encoders for long documents focuses on monolingual (mainly English) models. The few multilingual exceptions (Yu et al., 2021; Sagen, 2021) derive a multilingual Longformer from standard MMTs (XLM-R and mBERT) in exactly the same fashion in which the original work (Beltagy et al., 2020) pretrains English Longformer after initialization from RoBERTa weights. In this work, we replicated this effort, evaluating mLongformer as the main baseline for HMDE.

Pretraining for Retrieval. Self-supervised and distantly-supervised approaches have recently been proposed for pretraining documents encoders specifically for the task of document retrieval (Izacard et al., 2022; Yu et al., 2021; Gao et al., 2022). Izacard et al. (2022) pretrain Contriever – a BERT-based document encoder with an objective based on the inverse cloze task (Lee et al., 2019): a positive query-document pair is created by extracting a span of text from the document and using it as a "query"; they train with a contrastive objective that scores the document from which the query was extracted higher than other documents. Gao et al. (2022) feed queries as prompts to a generative language model, which then generates document; they then use Contriever to embed this synthetic document and find most similar real documents in the collection, finally fine-tuning Contriever on query-document pairs obtained this way. In a manner similar to ours, Yu et al. (2021) leverage Wikipedia as a source of quasi-parallel data: while we exploit document-level alignments, they leverage section-level aligments to create positive cross-lingual training instances for paragraph retrieval: a section title ("query") in one language is coupled with the section body ("document") in another language; they then train a multilingual Longformer initialized from mBERT with a combination of query MLM-ing and contrastive relevance ranking. In contrast to these efforts, we create a general-purpose (i.e., task-agnostic) multilingual document encoder that can both be fine-tuned for supervised tasks and used as a standalone document embedder.

Mining Parallel Documents. Mining parallel documents – a task which aims to identify mutual translations in a large document collection and is often used as a first step in extracting parallel sentences (Resnik and Smith, 2003; Uszkoreit et al., 2010; Schwenk, 2018, *inter alia*) – is the task that bears most resemblance to our pretraining. Transformer-based approaches to the task (Guo et al., 2019; El-Kishky and Guzmán, 2020; Gong et al., 2021) typically aggregate document-level representations from multilingual sentence embeddings. The work of Guo et al. (2019) is arguably most related to ours: they train a hierarchical encoder with a simple feed-forward net as the upper encoder that independently transforms precomputed sentence embeddings: document embedding is then the average of feed-forward-transformed sentence embeddings. The model is trained bilingually (English-Spanish and English-French) with a contrastive objective on a huge silver-standard corpus of parallel documents (13M and 6M document pairs, respectively) and evaluated on the very same task of parallel document mining. Our work differs in two crucial aspects: (1) while (Guo et al., 2019) train *bilingual* models for recognizing parallel documents, we train a single general-purpose massively multilingual document encoder; (2) we train on a much smaller corpus of comparable (not parallel) documents, readily available from Wikipedia. Both aspects make HMDE much more widely applicable, for both supervised and unsupervised document-level tasks and any of the languages from LaBSE's pretraining (as HMDE's lower encoder is initialized with LaBSE's weights).

6 Conclusion

In this work, we pretrain a multilingual document encoder based on a hierarchical transformer architecture (HMDE), and initialize its lower-level encoder with the weights of a state-of-the-art multilingual sentence encoder. We leverage Wikipedia as a rich source of quasi-parallel long documents and train HDME with a contrastive cross-lingual document matching objective. We show that the obtained model is a general-purpose multilingual document encoder that can successfully be both (1) fine-tuned for document-level cross-lingual transfer and (2) used as a document embedding model out of the box. Our results render HMDE substantially more effective than both multilingual Longformer and segmentation-based document encoding. Crucially, HMDE generalizes well to languages unseen in its document-level pretraining. Our follow-up experiments reveal that the size of the pretraining corpus affects the performance more than the number and diversity of languages involved, suggesting that reliable massively multilingual document encoders do not necessarily require equally massively multilingual pretraining.

Limitations

Because we initialize the lower transformer of HMDE with LaBSE (Feng et al., 2022), the set of languages that HMDE "supports" out of the box is bound to the set of 109 languages included in LaBSE's pretraining.[10] This means that HMDE will, in principle, be less effective as a document encoder for other languages.[11] HDME, like LaBSE, should in principle be useless for languages written in a script that LaBSE (or in fact, mBERT, from which LaBSE borrows the vocabulary and pretrained subword embeddings) has not seen in its pretraining, as the corresponding tokenizer will produce a sequence of unknown tokens ([UNK]). This means that HMDE, much like the rest of existing multilingual encoders, supports only a small fraction of world's 7000+ languages (Joshi et al., 2020). Moreover, all languages included in our evaluation datasets – MLDOC and CLEF – are covered by this set of 109 languages, which means that the average performance we report is likely a gross overestimate for languages unseen in LaBSE's pretraining. Further, HMDE leverages Wikipedia for training (with sets of either 4 or 12 languages, see 3.1) – the number of Wikipedia pages (and more generally, digital footprint of a language on the web) varies tremendously across languages, effectively limiting the selection of languages for HMDE's document-level pretraining. Our results (see 4.1), however, show that HMDE generalizes well to languages not seen in its document-level pretraining.

Further, HMDE is implemented as a Bi-Encoder (aka Siamese network), which means that for a given pair of documents in a training example (positive or negative pair), it separately encodes each of the documents. Cross-Encoder architecture, in which the documents would be concatenated before encoding, would have the advantage of allowing the encoder to contextualize the token/sentence representations of one document with those of the other before the computation of their similarity score. Cross-encoding architectures have been shown effective, albeit not efficient (i.e., slow) in training for document retrieval, in which the (short) query is concatenated with the (long) document (MacAvaney et al., 2020; Shi et al., 2020; Rosa et al., 2022). We do not explore cross-encoding in our work; in our case, it implies joint encoding of the concatenation of two long documents (in different languages), arguably exploding in GPU memory occupancy and possibly preventing us from fitting even single-instance batches on our GPU cards.

Ethical Considerations

We do not test HMDE explicitly to check whether the representations it produces reflect negative societal biases and stereotypes (e.g., sexism or racism), but given that its lower encoder is initialized from LaBSE's weights, it would not be surprising if this was the case. If so, many of the existing techniques from the literature designed to debias pretrained language models (Qian et al., 2019; Barikeri et al., 2021; Guo et al., 2022) could be applied to HMDE too, and in principle "as-is" (i.e., without special modifications).

References

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. Mad-g: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955.

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

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022. An exploration of hierarchical attention transformers for efficient long document classification. *arXiv preprint arXiv:2210.05529*.

[10] The full list is provided in Table 10 of the Appendix in (Feng et al., 2022).

[11] Not necessarily the case only for unseen that are close relatives to some of the high-resource languages seen in LaBSE's pretraining.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

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)*, pages 4171–4186.

SiYu Ding, Junyuan Shang, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-Doc: A retrospective long-document modeling transformer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2914–2927, Online. Association for Computational Linguistics.

Xin Dong, Yaxin Zhu, Yupeng Zhang, Zuohui Fu, Dongkuan Xu, Sen Yang, and Gerard de Melo. 2020. Leveraging adversarial training in self-learning for cross-lingual text classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1541–1544, New York, NY, USA. Association for Computing Machinery.

Ahmed El-Kishky and Francisco Guzmán. 2020. Massively multilingual document alignment with cross-lingual sentence-mover's distance. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 616–625.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721.

Goran Glavaš and Swapna Somasundaran. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7797–7804.

Hongyu Gong, Vishrav Chaudhary, Yuqing Tang, and Francisco Guzmán. 2021. Lawdr: Language-agnostic weighted document representations from pre-trained models. *arXiv preprint arXiv:2106.03379*.

Mandy Guo, Yinfei Yang, Keith Stevens, Daniel Cer, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Hierarchical document encoder for parallel corpus mining. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 64–72.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 8:1–22.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183.

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*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sean MacAvaney, Luca Soldaini, and Nazli Goharian. 2020. Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. In *European Conference on Information Retrieval*, pages 246–254. Springer.

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*, pages 1155–1156.

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

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1015–1025.

Magdalena Plamadă and Martin Volk. 2013. Mining for domain-specific parallel text from wikipedia. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 112–120.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228.

Philip Resnik and Noah A Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Guilherme Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. In defense of cross-encoders for zero-shot retrieval. *arXiv preprint arXiv:2212.06121*.

Markus Sagen. 2021. Large-context question answering with cross-lingual transfer. Master's thesis, Uppsala University, Department of Information Technology.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Peng Shi, He Bai, and Jimmy Lin. 2020. Cross-lingual training of neural models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2768–2773, Online. Association for Computational Linguistics.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. 2020. Sparse sinkhorn attention. In *International Conference on Machine Learning*, pages 9438–9447. PMLR.

Jakob Uszkoreit, Jay M Ponte, Ashok C Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1101–1109.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1725–1734.

Puxuan Yu, Hongliang Fei, and Ping Li. 2021. Cross-lingual language model pretraining for retrieval. In *Proceedings of the Web Conference 2021*, pages 1029–1039.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

A Appendix

A.1 Training and Optimization Details

In all training procedures, we use AdamW (Loshchilov and Hutter, 2019) as the optimization algorithm.

HMDE Pretraining. We set the maximal sentence length for HMDE, input to its lower-level transformer (initialized with LaBSE weights) to 128 tokens. For fair comparison, we set the segment size of the LasBSE-Seg baseline also to $N_S = 128$ tokens. For fair comparison against mLongformer, we limit the maximal document length for HMDE to 32 sentences, not to exceed the mLongformer's maximal input length of $4,096$ tokens. In our main set of experiments, the document-level (upper) transformer consists of 2 transformer layers, with GELU activation (Hendrycks and Gimpel, 2016), layer normalization ($\epsilon = 1\mathrm{e}^{-12}$), and feed-forward sublayer with hidden size of 2048. The dropout rate for the upper transformer is set to 0.1. We train in batches of size $N = 2$ with the gradient accumulation over 64 batches for 1 full epoch,[12] with the initial learning rate of $1\mathrm{e}^{-5}$, linear scheduling and 1000 warm-up steps.

mLongformer Pretraining. We train the mLongformer model (also initialized from LaBSE), also for 1 full epoch via MLM-ing, masking out 15% of tokens. We train with the initial learning rate of $1\mathrm{e}^{-5}$ with weight decay of 0.01 and 500 warm-up steps. We train in batches of size 2, accumulating gradients over 32 batches.

XLDC Fine-Tuning. We fine-tune both HMDE and mLongformer for topical document classification with the learning rate of $2\mathrm{e}^{-5}$ and without weight decay (with a 200 warm-up steps). We train in batches of size 4 for 50 epochs, accumulating gradients over 8 batches. Model selection was carried out based on the performance on the English validation portion of the MLDOC dataset, with early stopping if validation loss did not improve over 7 epochs.

A.2 Additional Ablation

We additionally test our design decision to segment the document into sentences, and encode sentences with the lower-level transformer (the weights of which are initialized from LaBSE). To this end, we compare our default strategy of segmenting input documents into sentences against a less-informed segmentation into consecutive chunks of 128 tokens. Table 4 shows the results of this comparison. Unsurprisingly – given that the lower encoder is initialized with the weights of a pretrained *sentence* encoder – sentence-based segmentation is more effective, although chunking does not trail by much.

Model	Segmentation	XLDC	CLIR
HMDE-LaBSE	*Sentence*	86.8	0.249
HMDE-LaBSE	*Chunk*	85.4	0.224

Table 4: HMDE results for different choices w.r.t. to document segmentation. Training for both variants carried out on XLW-4L. Results are averages across all test languages (XLDC) and language pairs (CLIR).

[12] Note that batch size $N = 2$ in our contrastive objective (see §2.2) implies only one in-batch negative pair (besides the hard negative) for each positive pair.