Unsupervised Context Aware Sentence Representation Pretraining for Multi-lingual Dense Retrieval

Ning Wu[1] , **Yaobo Liang**[2] , **Houxing Ren**[1] , **Linjun Shou**[1] , **Nan Duan**[2] , **Ming Gong**[1] and **Daxin Jiang**[1]

[1]Microsoft STCA
[2]Microsoft Research Asia
{wuning, yalia, v-houxingren, lisho, nanduan, migon, djiang}@microsoft.com

Abstract

Recent research demonstrates the effectiveness of using pretrained language models (PLM) to improve dense retrieval and multilingual dense retrieval. In this work, we present a simple but effective monolingual pretraining task called contrastive context prediction (CCP) to learn sentence representation by modeling sentence level contextual relation. By pushing the embedding of sentences in a local context closer and pushing random negative samples away, different languages could form isomorphic structure, then sentence pairs in two different languages will be automatically aligned. Our experiments show that model collapse and information leakage are very easy to happen during contrastive training of language model, but language-specific memory bank and asymmetric batch normalization operation play an essential role in preventing collapsing and information leakage, respectively. Besides, a post-processing for sentence embedding is also very effective to achieve better retrieval performance. On the multilingual sentence retrieval task Tatoeba, our model achieves new SOTA results among methods without using bilingual data. Our model also shows larger gain on Tatoeba when transferring between non-English pairs. On two multi-lingual query-passage retrieval tasks, XOR Retrieve and Mr.TYDI, our model even achieves two SOTA results in both zero-shot and supervised setting among all pretraining models using bilingual data.

1 Introduction

Nowadays, cross-lingual pre-training [Devlin *et al.*, 2019; Conneau and Lample, 2019; Huang *et al.*, 2019; Conneau *et al.*, 2020] has achieved great performance on cross-lingual transfer learning tasks. These pre-trained models could fine-tune on one language and directly test on other languages. Cross-lingual sentence representation models like InfoXLM [Chi *et al.*, 2021] and LaBSE [Feng *et al.*, 2020] target to generate good cross-lingual representation without fine-tuning. These models use contrastive loss to make bilin-

gual sentence pairs have similar embeddings and achieve great performance on bilingual sentence retrieval tasks.

However, there are two potential problems for these methods. First, most of them rely on bilingual corpus, which are not always available, especially for low-resource languages and non-English languages pairs. Only using English related bilingual pairs will limit the transferability between non-English pairs. Second, multilingual dense retrieval tasks such as XOR Retrieve and Mr.TYDI require the model to map semantic related query and passage to similar position in embedding space, but existed methods only could map bilingual sentence pairs with same meaning to similar embedding.

(a) XLM-R.

(b) XLM-R with Calibration.

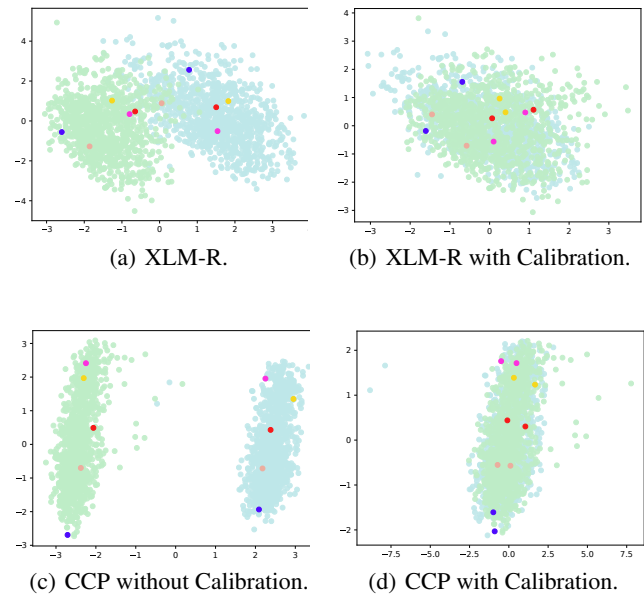(c) CCP without Calibration.

(d) CCP with Calibration.

Figure 1: Visualization of the representation output of four models. A thousand of bilingual sentence pairs from Tatoeba are converted into representations by four models. The dimension of these representations is reduced from 768 to 2 by Principle Component Analysis (PCA), so that a sentence can be mapped into a point on figure (a), (b), (c) and (d). For each figure, we highlight five bilingual pairs, and the two highlighted cycles in the same color denote a bilingual sentence pair from English and Arabic, respectively.

Inspired by recent progress of contrastive learning on dense

retrieve, we propose a new pretraining task called Contrastive Context Preiction (CCP). CCP targets on constructing isomorphic embedding space by modeling the sentence level contextual relation in long documents. Formally, a document is a sequence of sentences. For each center sentence, we define the sentence in the window centered on it as a context sentence. First, CCP will encode each sentence into a vector. Given embedding of center sentence s_c, the model need to select correct context sentence s out of thousands random sampled sentences and vice versa. With contrastive context prediction, we could estimate the mutual information $I(s|s_c)$ of contextual relation. Our experiments show that the embeddings of CCP have isomorphic structure across different languages. Furthermore, we do cross-lingual calibration to further improve the alignment. We illustrate our ideas with Figure 1, which visualizes the English and Arabic embedding space from XLM-R and our models. We randomly highlight five points for each languages and the sentences with same meaning have same color. The distribution of two languages shown in Figure 1(a) doesn't have an obvious pattern. With contrastive context prediction, the five points show similar relative position and the shape of all points are similar to each other, but the embeddings of two languages are spread in two regions of latent space. After cross-lingual calibration, the points with same color almost have similar position.

We evaluate our model on bilingual sentence retrieval task Tatoeba, which can test whether our model could generate similar embeddings for two sentences with same meaning but from different languages. Our model achieves SOTA results among methods without bilingual data, and our results are very close to the model with bilingual data. Besides, our model shows better cross-lingual transferability between non-English languages pairs. On two multi-lingual query-passage retrieval tasks, XOR Retrieve and Mr.TYDI, CCP achieves new SOTA results among all pretraining models using bilingual data.

Our contribution can be summarized as:

- We propose a contrastive context prediction pretraining (CCP) task that is capable of learning isomorphic representations for each language without parallel data. Our model can achieve excellent performance on multilingual retrieval, especially between two non English-centric languages-pairs.

- We design an effective contrastive pretraining framework for sentence embedding pretraining. It consists of language-specific memory bank and projection head with asymmetric batch normalization. Both of them play essential role in preventing collapsing. And we also observe the offset phenomenon on on bilingual sentence representation pair produced by our model. To align the sentence embedding space between different languages, we propose cross-lingual calibration to align the bilingual sentence pair into the same position in latent space.

- We conduct evaluation experiments upon three types of multi-lingual retrieval tasks. Extensive results on the three datasets have shown superiority of the proposed model in both effectiveness and rubustness.

2 Related Work

Cross-lingual Pre-trained Model Multilingual BERT (M-BERT) [Devlin *et al.*, 2019] performs pre-training based on the multilingual corpus with the masked language model task. By sharing the model parameters and the vocabulary across all languages, M-BERT obtains the cross-lingual capability over 102 languages. XLM [Conneau and Lample, 2019] performs cross-lingual pre-training based on multilingual corpus and bilingual corpus, by introducing the translation language model task into pre-training. Based on XLM, Unicoder [Huang *et al.*, 2019] uses more cross-lingual pre-training tasks and achieves better results on XNLI. XLM-R [Conneau *et al.*, 2020] is a RoBERTa [Liu *et al.*, 2019]-version XLM without using translation language model in pre-training. It is trained based on a much larger multilingual corpus (i.e. Common Crawl) and becomes the new state-of-the-art. Both these models and our model could achieve good performance after fine-tuning. Our model also could produce good cross-lingual sentence embedding without fine-tuning.

Dense Passage Retrieval [Lee *et al.*, 2019] proposed a simple Inverse Cloze Task (ICT) method to further continue-train BERT. REALM[Guu *et al.*, 2020] is an end-to end co-training framework for reader and retriever. karpukhin-etal-2020-dense is the first to discover that careful fine-tuning can learn effective dense retriever directly from BERT. Later works then started to investigate ways to further improve fine-tuning. ANCE [Xiong *et al.*, 2020] selects hard training negatives globally from the entire corpus, using an asynchronously updated ANN index. [Qu *et al.*, 2021] proposed the RocketQA fine-tuning pipeline which hugely advanced the performance of dense retrievers. coCondenser [Gao and Callan, 2021] is one of the best dense passage retrieval model on MS-MARCO, Natural Question and Trivia QA. It adds an unsupervised corpus-level contrastive loss to warm up the passage embedding space.

Contrastive Learning CPC [Oord *et al.*, 2018] predicts the future in latent space by using powerful auto-regressive models. It uses a probabilistic contrastive loss which induces the latent space to capture information that is maximally useful to predict future samples. [Wu *et al.*, 2018] presents an unsupervised feature learning approach by maximizing distinction between instances via a novel non-parametric softmax formulation, which is so-called memory bank mechanism. SimCLR [Chen *et al.*, 2020] simplifies recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. MoCo [He *et al.*, 2020] maintains a dynamic dictionaries by implementing a momentum-based moving average mechanism of the query encoder. InfoXLM [Chi *et al.*, 2021] contains a pre-training task based on contrastive learning. Given a bilingual sentence pair, they regard them as two views of the same meaning, and encourage their encoded representations to be more similar than the negative examples. LaBSE[Feng *et al.*, 2020] combines masked language model (MLM) and translation language model (TLM)[Conneau and Lample, 2019] pretraining with a translation ranking task using bi-directional dual encoders. CLEAR [Wu *et al.*, 2020] performs contrastive learning on for sentence representation. It employs multiple
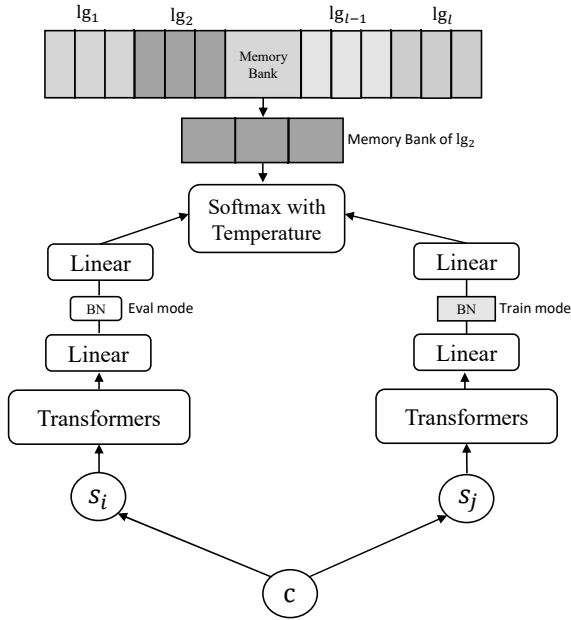
Figure 2: Overview of our framework of contrastive context prediction, the proposed representation learning framework. Two contextual sentences are encoded by training mode batch normalization and testing mode batch normalization, respectively. Finally, we add negative samples of the same language from memory bank in contrastive loss calculation.

sentence-level augmentation strategies in order to learn a noise-invariant sentence representation. We follow SimCLR on the detailed implementation of contrastive learning. Different from these works, we leverage contrastive learning to model the sentence-level contextual relation in natural languages.

3 Methodology

Our method contains two steps. First, we train the model with Masked Language Model and Contrastive Context Prediction. This step targets on building isotropic sentence embedding space for each language. The learned sentence embeddings from different languages will have good cross-lingual property after they became isomorphic by CCP task. Just as Figure 1(c) shows, different languages are spread in different region of latent space with similar structure. Thus in second step, we use cross-lingual calibration to further align different languages.

3.1 Contrastive Loss for Context Prediction

Our method targets to model the sentence-level contextual relation. Formally, a document D is a sequence of sentences $(s_1, s_2, ..., s_l)$. For each center sentence s_c, we define its contextual set as $Context(s_c) = \{s_p | c-w <= p <= c+w, p \neq c\}$, w is radius of window which represents the maximum distance between center sentence and context sentence. Our model will model the relation between center sentence and its contextual sentences. In this subsection, we introduce the details of contrastive loss $L_N(s_i, s_c)$.

Scoring Function The scoring function $f(s_i, s_c)$ takes two sentences as input and output a score. To begin with, we encode s and s_c to a vector separately. In this step, we choose

a Transformer-based encoder. We use the corresponding representation of a manually-inserted token [CLS] as the embedding of the whole sentence. Then, we add a non-linear neural network, namely Projection Head, to further map the vector to a new space. The projection head consists of two linear layers and one batch normalization between two layers. We denote the representation of s_i and s_c as z_i and z_c respectively. Following SimCLR [Chen *et al.*, 2020], we only use it when computing contrastive loss and abandon it after pre-training. Projection Head could help model learn general representations and will not overfit to contrastive loss. Finally, we choose scoring function as $f(s_i, s_c) = exp(cosine(z_i, z_c)/\tau)$, where τ represents temperature and is a hyper-parameter.

Language-specific Memory Bank In [Oord *et al.*, 2018], the lower bound of contrastive loss becomes tighter as negative samples number N becomes larger. To further increase N while batch size is limited by GPU memory, we use memory bank to store the embeddings from recent batches and use them in the training of current batch. Because we are handling multiple languages at same time, we tried two strategies: plain memory bank and language-specific memory bank, and find later one is better. The "language-specific" means that our memory bank will tag the embeddings by language. For each language, it will only use the embeddings from the same language in training. We tried to use shared memory bank for all languages. Since a shared memory bank provides negative samples from various languages, the model will focus on classifying the language of sentences, rather than learning contextual relation. Hence the loss could be very small because language classification is very easy, but the cross-lingual performance will be very poor. The memory bank is maintained in FIFO (First-In-First-Out) manner. During each iteration, the representation z_i as well as the network parameters θ are optimized via Adam. Then z_i is added to M, and the oldest representation in the memory bank is deleted.

Asymmetric Batch Normalization Batch normalization is an essential part of projection head. However, traditional batch normalization is easy to be trapped in information leak problem [He *et al.*, 2020], which means that the contrastive loss is very small but the evaluation results on down stream tasks are very low. Hence, we propose asymmetric batch normalization to avoid information leak. It is more efficient than shuffle batch normalization [He *et al.*, 2020], which requires communication between GPUs. In training procedure, the mode of batch normalization will be changed in training mode and testing mode alternately and the mode of batch normalization in two projection heads should be kept different in training. In testing mode, the batch normalization will use running mean and running variance to replace batch mean and batch variance, which is able to prevent information leak. The detail training procedure can be seen in Algorithm 1. Compared with shuffle batch normalization proposed in [He *et al.*, 2020], our method is very easy to implement and has good performance.

3.2 Cross-lingual Calibration

After we acquire isomorphic sentence embeddings by pretraining, in order to better align between languages, we further propose cross-lingual calibration for sentence representation.

Algorithm 1 The training algorithm for the contrastive context prediction task.

1: **Input:** Batch size N, constant τ, structure of f, g.
2: **Output:** Model parameters Θ.
2: **for** sampled minibatch $\{c_k\}_W^N$ **do**
3:    draw center sentence s_c from context c_w.
4:    randomly draw target sentence s_i from context c_w.
5:    $h_c = f(s_c); h_i = f(s_i);$
6:    **if** training mode $e = 0$ **then**
7:       $g_c() = g().\text{train}(); g_i() = g().\text{eval}();$
8:       $z_c = g_c(h_c); z_i = g_i(h_i); e = 1;$
9:    **else**
10:      $g_c() = g().\text{eval}(); g_i() = g().\text{train}();$
11:      $z_c = g_c(h_c); z_i = g_i(h_i); e = 0;$
12:    **end if**
13:    $l_{c,i}^w = -\log \frac{\exp(\cos(z_c, z_i)/\tau)}{\sum_{k=1}^{2N + M_{lg(i)}} \mathbb{1}_{[k \neq c]} \exp(\cos(z_c, z_k)/\tau)}$
14:    $\mathcal{L}_{CL} = \sum_{c=1}^{2N} \sum_{i=1}^{2N} m(s_c, s_i) l_{c,i}^w$
15:    $m(s_c, s_i) = 1$ means c and i exists in the same local window, unless $m(s_c, s_i) = 0$.
16:    update networks f and g to minimize $\mathcal{L}_{CL}$.
16: **end for**
17: **return** encoder network $f(\cdot)$, and throw away $g(\cdot) = 0$

The cross-lingual calibration consists of three operations: shifting, scaling and rotating. We do three operations separately to better understand the properties of latent space. For shifting, we compute the mean of sentence embedding μ from different languages. Then we acquire shifted mean sentence representation by subtracting corresponding language mean vector. For Scaling, we compute the variance of sentence embedding σ for all languages based on Common Crawler corpus. For a shifted sentence, we acquire a scaled sentence representation by dividing corresponding language variance vector. Fianlly, we learn the rotation matrix in unsupervised method proposed by [Conneau *et al.*, 2018]. With orthogonal rotation matrix $W_{i,j}$, sentence embedding of $lg(i)$ can be mapped to $lg(j)$. We provide detailed description of this part in the supplementary materials.

4 Experiments

In this section, we first set up the experiments, and then present the performance comparison and result analysis.

Our CCP model has 1024 hidden units, 16 attention heads and 24 layers in encoder. Following [Wenzek *et al.*, 2019], we collect a clean version of Common Crawl as pre-training corpus. It leads to 2,500GB multilingual corpus covering 108 languages. We first initialize the CCP model with XLM-R[Conneau *et al.*, 2020], and then run continued pre-training with the accumulated 2,048 batch size with gradient accumulation and a memory bank of 32768. One 2,048 batch consists of many small batches whose size is 32 for CCP. We use Adam Optimizer with a linear warm-up and set the learning rate to 3e-5. We select two pre-training tasks randomly in different batches. This costs 7 days on 16 V100 for CCP model.

4.1 Baselines

Here are the baselines for our experiments.

  •*M-BERT* [Devlin *et al.*, 2019] is a multilingual version of BERT.

  •*XLM-R* [Conneau *et al.*, 2020] uses a Transformer-based masked language model on one hundred languages.

  •*InfoXLM* [Chi *et al.*, 2021] formulates a cross-lingual pre-training as maximizing mutual information between multilingual multi-granularity text.

  •*Unicoder* [Liang *et al.*, 2020] uses mask language model and translation language model as pre-training tasks.

  •*CRISS* [Tran *et al.*, 2020] utilizes cross-lingual retrieval for iterative training.

  •*LaBSE* [Feng *et al.*, 2020] formulates a translation ranking task using bi-directional dual encoders.

We present the comparison between our method and AB-Sent [Fu *et al.*, 2020] in appendix, since it only reports their performance on 3 language pairs of tatoeba, All transformer models in this paper use Bert-large structure except CRISS and LaBSE. CRISS follows mBART structure with 24 layers transformer, and LaBSE follows Bert-base structure with a 500k size vocabulary, which is twice as large as our model.

4.2 Cross-lingual Sentence Retrieval

To better evaluate the performance on massive languages, we adopt the Tatoeba corpus introduced by [Artetxe and Schwenk, 2019]. It consists of 1,000 English-centric sentence pairs for 112 languages and the task aims to find the nearest neighbor for each sentence in the other language using cosine similarity distance. To compare with previous model, we only report results on 14 language in experiments, and we present all results in Table 2 of supplementary material. Besides the English-centric dataset constructed by [Artetxe and Schwenk, 2019], we choose 14 language-pairs which don't contain English in the Tatoeba raw dataset, and we present results on 50 language-pairs which don't contain English in Table 2 of supplementary material. Following [Artetxe and Schwenk, 2019], we extract 1000 sentence-pairs for each language pair and test the pre-trained model on the Tatoeba dataset without fine-tuning directly. The accuracy for each language pair is computed. We report CCP in Table 1 and Table 2. For large models, we use the the averaged hidden vectors in the 14-th layers as sentence representation for sentence retrieval.

In Table 1, (1) we find CCP performs significantly better than XLM-R and CRISS and achieves new SOTA results among methods without using bilingual data on the Tatoeba dataset. (2) Compared with models using bilingual data, CCP performs better than InfoXLM [Chi *et al.*, 2021], Unicoder [Liang *et al.*, 2020] and LASER [Artetxe and Schwenk, 2019], but it is worse than LaBSE [Feng *et al.*, 2020], which is the SOTA model with bilingual data on Tatoeba. LaBSE uses bidirectional dual encoders with 8192 batch size to learn cross-lingual sentence representation. The translation ranking task is similar to contrastive learning task, which is severely affected by the batch size. Limited by hardware, we can only perform contrastive learning with a batch size of 32 , it's hopeful our model will have a better result with larger batch.

In Table 2, we find (1) CCP performs significantly better than XLM-R and CRISS. We observe that the performance of CRISS decreases obviously, because CRISS only mines

Type	Methods	FR	ES	DE	EL	BG	RU	TR	AR	VI	TH	ZH	HI	SW	UR	AVG	AVG$_{all}$
M	XLM-R	56.0	56.7	72.0	35.4	48.5	50.0	49.6	31.5	45.9	36.5	42.1	47.8	11.0	32.0	43.9	-
	CRISS	92.7	96.3	98.0	-	-	90.3	92.9	78.0	92.8	-	85.6	92.2	-	-	-	-
B+M	INFOXLM	83.7	87.8	94.7	67.1	78.9	84.9	83.5	63.5	89.8	86.7	84.9	86.4	35.8	69.4	78.4	-
	LASER	95.7	98.0	97.3	95.0	95.1	94.6	97.6	91.9	96.8	95.4	95.5	94.7	57.6	81.9	91.9	65.5
	Unicoder	81.6	86.5	93.8	67.2	77.5	81.6	76.7	53.4	80.9	70.2	87.7	73.8	30.3	59.2	72.9	-
	LABSE	**96.0**	**98.4**	**99.4**	**96.6**	**95.7**	**95.3**	**98.4**	**91.0**	**97.8**	**97.1**	**96.2**	**97.7**	**88.6**	**95.4**	**96.0**	**83.7**
M	CCP	93.8	96.6	98.5	87.6	88.2	92.0	95.2	81.5	94.3	90.2	91.8	90.4	50.5	82.8	88.1	-
	CCP+Calibration	94.9	97.2	99.0	93.0	90.3	93.5	97.1	87.9	96.3	95.3	95.0	96.2	64.2	91.3	92.2	78.8

Table 1: Evaluation results on English-centric cross-lingual sentence retrieval. Type means if a model uses bilingual data (B) and monolingual (M) data in pre-training. Given each model, the corresponding retrieval results on all languages are listed in the same row. We report the average Top-1 accuracy of two direction(e.g. EN-FR and FR-EN). AVG$_{all}$ is the average of 112 languages that Tatoeba supports.

Type	Methods	DE-EL	DE-IT	RU-NL	FR-DE	IT-RU	AR-RU	ZH-ES	ZH-JA	JA-FR	ES-PT	IT-RO	SV-DA	DA-NO	NL-DE	UR-RU	AVG
M	XLM-R	54.6	54.9	66.6	75.8	47.4	43.4	47.0	52.8	42.5	76.0	51.4	81.1	89.4	66.7	85.6	62.3
	CRISS	-	77.5	83.5	88.4	74.7	79.6	80.7	72.5	74.0	-	-	-	-	-	-	-
B+M	Unicoder	68.5	68.5	77.3	81.9	64.7	64.5	65.9	67.3	55.3	82.4	61.9	85.4	91.9	72.4	90.1	73.2
	LABSE	**85.6**	**80.3**	**89.9**	**90.4**	**78.1**	**87.3**	**88.5**	**88.6**	**89.1**	**84.3**	**72.6**	87.6	**93.5**	78.1	91.7	**85.7**
M	CCP	83.5	78.6	87.5	89.0	76.3	83.6	83.9	84.1	82.6	84.0	70.7	87.7	93.1	78.2	92.1	83.7
	CCP+Calibration	85.2	80.0	89.1	89.7	77.8	86.2	87.4	86.5	87.6	83.7	71.2	**87.9**	93.2	**78.4**	**92.3**	85.1

Table 2: Evaluation results on Non-English cross-lingual sentence retrieval. Type means if a model uses bilingual data (B) and monolingual (M) data in pre-training. Given each model, the corresponding retrieval results on all languages are listed in the same row. We report the average Top-1 accuracy of two direction(e.g. DE-EL and EL-DE).

English-centric bilingual pairs and it is easy to overfit English-centric sentence retrieval. (2) For models using bilingual corpus, CCP performs better than Unicoder and its performance is very close to LaBSE. Since LaBSE is trained on English-centric bilingual corpus, its performance decreases severely on sentence retrieval between two non-English sentences, which means it is overfitting on English-centric sentence retrieval. Compared with these English-centric models (LABSE, Unicoder, InfoXLM, CRISS) , our model is more general, and it does not rely on English data in down-streaming tasks. So our model has lesser performance loss when it is evaluated on non English-centric sentence retrieval.

These two experiments proved that CCP has good cross-lingual sentence retrieval performance both on English-centric and Non-English languages pairs. Our performance is better than all models without using bilingual data and slightly worse than LaBSE which used bilingual data.

4.3 Cross Lingual Query Passage Retrieval

We further evaluate cross lingual transfer-ability on two new zero-shot settings: 1. Give a query from language L, retrieve relevant passages which can answer the query from the English corpus. 2. Give a query from language L, retrieve relevant passages which can answer the query from language L.

Hence, we adopt XOR-QA [Asai *et al.*, 2020] dataset and Mr. TYDI [Zhang *et al.*, 2021] dataset to evaluate our method on the two settings. Both of the two datasets are constructed from TYDI, a question answering dataset covering eleven typologically diverse languages. The XOR-QA dataset consists of three tasks: XOR-Retrieve, XOR-English Span, and XOR-Full. We take the XOR-Retrieve task to evaluate our method. XOR-Retrieve is a cross-lingual retrieval task where the query is written in a target language (e.g., Japanese) and the model is required to retrieve English passages that can answer the query. Same to the source paper [Asai *et al.*, 2020], we measure the recall by computing the fraction of the questions for which

the minimal answer is contained in the top n tokens selected. We evaluate with n = 2k, 5k: R@2kt and R@5kt (kilo-tokens). The Mr. TYDI dataset is a multi-lingual benchmark dataset for mono-lingual query passage retrieval in eleven typologically diverse languages, designed to evaluate ranking with learned dense representations. Same to the source paper [Zhang *et al.*, 2021], we use MRR@100 and Recall@100 as metrics.

In this paper, we adopt a zero-shot setting to evaluate our method. We train the pre-trained model on Natural Question data and directly test the model on XOR-QA [Asai *et al.*, 2020] dataset and Mr. TYDI [Zhang *et al.*, 2021] dataset. We train the model on 8 NVIDIA Tesla V100 GPUs (with 32GB RAM). We use AdamW Optimizer with a learning rate of 1e-5. The model is trained up to 20 epochs with a mini-batch size of 48. The rest hype-parameters are the same as DPR [Karpukhin *et al.*, 2020]. Besides, we also report our supervised result on XOR-Retrieve in appendix. We achieve 2% advantage over the second-placed model. Besides, we also experiment our model on XOR Retrieve in supervised setting with the same hyperparameter settings, and achieve SOTA results with 2% advantage over the second-placed model on leaderboard[1].

In Table 3, (1) we find CCP performs significantly better than XLM-R , InfoXLM and LaBSE. CCP achieves new SOTA results among methods without using bilingual data on the Mr. Tydi dataset. (2) Especially, we find InfoXLM has good performance on the 15 languages that it has bilingual corpus. However, on the languages that it doesn't have bilingual corpus, such as BN, FI, and TE, we find it's worse than LaBSE. For CCP, we find it has best performance on almost all languages, because our model is able to support 108 languages without collecting any bilingual corpus. In Table 4, we find (1) CCP performs slightly better than InfoXLM and have comparable performance on LaBSE. On these low-resource languages that InfoXLM can't cover, our method still have a

---

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

Methods	Metrics	AR	BN	EN	FI	ID	JA	KO	RU	SW	TE	TH	AVG
MBERT	MPR@100	0.301	0.303	0.283	0.226	0.319	0.243	0.211	0.267	0.185	0.120	0.174	0.239
	Recall@100	0.695	0.712	0.749	0.645	0.739	0.662	0.565	0.674	0.537	0.433	0.529	0.631
XLMR	MPR@100	0.365	0.374	0.275	0.318	0.395	0.299	0.304	0.306	0.274	0.346	0.401	0.333
	Recall@100	0.813	0.842	0.776	0.782	0.886	0.785	0.727	0.774	0.633	0.875	0.882	0.798
InfoXLM	MPR@100	0.373	0.354	0.325	0.300	0.380	0.310	0.299	0.313	0.351	0.311	0.400	0.338
	Recall@100	0.806	0.860	0.804	0.749	0.869	0.788	0.717	0.767	0.724	0.867	0.874	0.802
LABSE	MPR@100	0.372	**0.504**	0.314	0.309	0.376	0.271	0.309	0.325	**0.394**	0.465	0.374	0.365
	Recall@100	0.762	**0.910**	0.783	0.760	0.852	0.669	0.644	0.744	0.750	**0.889**	0.834	0.782
CCP	MPR@100	**0.426**	0.457	**0.359**	**0.372**	**0.462**	**0.377**	**0.346**	**0.360**	0.392	**0.470**	**0.489**	**0.410**
	Recall@100	**0.820**	0.883	**0.801**	**0.787**	**0.875**	**0.800**	**0.732**	**0.772**	**0.751**	0.888	**0.889**	**0.818**

Table 3: Evaluation results on Mr. TYDI. We use MPR@100 and Recall@100 as evaluation metrics. Given each model, the corresponding retrieval results on all languages are listed in the same row.

Methods	Metrics	AR	BN	FI	JA	KO	RU	TE	AVG
XLMR	Recall@2kt	0.414	0.470	0.529	0.407	0.439	0.566	0.639	0.452
	Recall@5kt	0.534	0.572	0.611	0.498	0.540	0.397	0.718	0.553
INFOXLM	Recall@2kt	0.485	0.520	0.516	0.407	**0.477**	0.325	0.668	0.485
	Recall@5kt	0.563	0.599	0.599	0.498	0.547	0.422	0.756	0.569
LABSE	Recall@2kt	0.469	**0.553**	0.487	0.382	0.418	0.333	**0.676**	0.474
	Recall@5kt	**0.566**	**0.661**	0.570	0.473	0.509	**0.439**	**0.790**	**0.573**
CCP	Recall@2kt	**0.472**	0.493	**0.570**	**0.423**	0.456	**0.342**	0.672	**0.490**
	Recall@5kt	0.553	0.586	**0.643**	**0.510**	**0.572**	0.414	0.777	0.570

Table 4: Evaluation results on XOR Retrieve. We use Recall@2kt and Recall@5kt as evaluation metrics. Given each model, the corresponding retrieval results on all languages are listed in the same row.

better performance. (2) We find our model has especially bad performance on Bengali (BN), which is consistent to results on Mr. TYDI. Our performance is better than all models using bilingual data, which means our context-aware pretraining is very suitable to various retrieval tasks, not only bilingual paraphrase retrieval, also query passage retrieval.

5 Ablation Study and Sensitivity Analysis

MB	L_2 norm	ABN	0.001	0.01	0.1	1.0
✓	✗	✗	fail	fail	fail	fail
✓	✓	✗	70.3	71.6	72.5	63.1
✓	✗	✓	63.1	61.3	65.4	58.1
✓	✓	✓	85.7	84.8	91.3	75.3
✗	✓	✓	81.7	82.8	90.3	72.1
✓	✓	BN	58.1	60.1	62.4	61.5

Table 5: Examining the influence of L_2 normalization , batch normalization, temperature and memory bank with sentence retrieval task between En and Fr.

Impacts of L_2 Normalization and Asymmetric Batch Normalization We next study the importance of L_2 normalization, batch normalization, and temperature τ in our contrastive loss. We use the pre-training and test setting in the last section. Table 5 shows that without L_2 normalization before softmax and batch normalization in projection head, our model will fail in training. In the last row, we present the result of a model with L_2 normalization and vanilla batch normalization, and it has a very terrible result. The model appears to "cheat" the pretext task and easily finds a low-loss solution. This is possibly because the intra-batch communication among samples (caused by BN) leaks information. [He *et al.*, 2020]. With asymmetric batch normalization, different mean and variance will be used to calculate z_c and z_i, respectively.

Windows Size / Batch Size	8	16	32	64
2	84.3	86.0	88.5	90.4
3	85.8	86.1	88.6	90.9
5	86.2	87.9	89.4	91.3

Table 6: Examining the influence of window size and batch size on the model performance between En and Fr of Tatoeba. Window size is 2 means predicting the next sentence.

Impacts of Batch Size and Window Size Our current model is highly dependent on context window and batch size affects the number of negative samples directly. We study the importance of batch size and window size in the pre-training stage. Since 108 languages pre-training is too time-consuming, we only use English and French as pre-training corpus, and report the sentence retrieval result between English and French. Table 6 shows that without large window and large batch, performance is slightly worse.

6 Conclusion

We propose a new cross-lingual pretrain task called Contrastive Context Prediction (CCP), and conduct comprehensive evaluations with interesting findings observed. We find CCP task is able to make sentence embedding space of different language isomorphic. The proposed approach achieves a excellent performance on multi-lingual dense retrieval.

References

[Artetxe and Schwenk, 2019] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, 7:597–610, 2019.

[Asai *et al.*, 2020] Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi.

Xor qa: Cross-lingual open-retrieval question answering. *arXiv preprint arXiv:2010.11856*, 2020.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020.

[Chi *et al.*, 2021] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *NAACL*, 2021.

[Conneau and Lample, 2019] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *NeuralIPS*, 2019.

[Conneau *et al.*, 2018] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *ICLR*, 2018.

[Conneau *et al.*, 2020] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *ACL*, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, June 2019.

[Feng *et al.*, 2020] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

[Fu *et al.*, 2020] Zuohui Fu, Yikun Xian, Shijie Geng, Yingqiang Ge, Yuting Wang, Xin Dong, Guang Wang, and Gerard de Melo. Absent: Cross-lingual sentence representation mapping with bidirectional gans. In *AAAI*, 2020.

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

[Guu *et al.*, 2020] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[Huang *et al.*, 2019] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *EMNLP*, 2019.

[Karpukhin *et al.*, 2020] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

[Lee *et al.*, 2019] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of ACL*, pages 6086–6096, 2019.

[Liang *et al.*, 2020] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *EMNLP*, pages 6008–6018, 2020.

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

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

[Qu *et al.*, 2021] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of NAACL-HLT*, 2021.

[Tran *et al.*, 2020] Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. Cross-lingual retrieval for iterative self-supervised training. *arXiv preprint arXiv:2006.09526*, 2020.

[Wenzek *et al.*, 2019] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzman, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

[Wu *et al.*, 2018] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[Wu *et al.*, 2020] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*, 2020.

[Xiong *et al.*, 2020] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

[Zhang *et al.*, 2021] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. tydi: A multi-lingual benchmark for dense retrieval. *arXiv preprint arXiv:2108.08787*, 2021.