Cross-lingual Contextualized Phrase Retrieval

Huayang Li[♣∗] **Deng Cai**[♡] **Zhi Qu**[♣] **Qu Cui**[♡]
Hidetaka Kamigaito[♣] **Lemao Liu**[♡] **Taro Watanabe**[♣]
[♣]Nara Institute of Science and Technology [♡]Tencent AI Lab
{li.huayang.lh6, qu.zhi.pv5, kamigaito.h, taro}@is.naist.jp
{jcykcai, qucui, redmondliu}@tencent.com

Abstract

Phrase-level dense retrieval has shown many appealing characteristics in downstream NLP tasks by leveraging the fine-grained information that phrases offer. In our work, we propose a new task formulation of dense retrieval, *cross-lingual contextualized phrase retrieval*, which aims to augment cross-lingual applications by addressing polysemy using context information. However, the lack of specific training data and models are the primary challenges to achieve our goal. As a result, we extract pairs of cross-lingual phrases using word alignment information automatically induced from parallel sentences. Subsequently, we train our **C**ross-lingual **C**ontextualized **P**hrase **R**etriever (CCPR) using contrastive learning, which encourages the hidden representations of phrases with similar contexts and semantics to align closely. Comprehensive experiments on both the cross-lingual phrase retrieval task and a downstream task, i.e, machine translation, demonstrate the effectiveness of CCPR. On the phrase retrieval task, CCPR surpasses baselines by a significant margin, achieving a top-1 accuracy that is at least 13 points higher. When utilizing CCPR to augment the large-language-model-based translator, it achieves average gains of 0.7 and 1.5 in BERTScore for translations from X⇒En and vice versa, respectively, on WMT16 dataset. We release our code and data at https://github.com/ghrua/ccpr_release.

1 Introduction

Compared with the dense retrieval at sentence (or passage) level (Karpukhin et al., 2020; Borgeaud et al., 2022; Asai et al., 2024), learning dense retrieval at the phrase level has shown more appealing characteristics in extensive NLP tasks, such as entity linking (Gillick et al., 2019), open-domain question answering (Lee et al., 2021a,b), text generation (Lan et al., 2023; Cao et al., 2024), etc. An important reason is that phrases can provide more fine-grained information than sentences.

In cross-lingual research, the phrase-level dense retrieval also shows the promise to solve a range of NLP problems (Bapna and Firat, 2019; Zheng et al., 2022). The cross-lingual phrase retrieval in Zheng et al. (2022) focuses on the mapping of cross-lingual Wiki entities, where each of them is represented by averaged hidden representations from 32 different contexts. However, the constrained phrase type, i.e., the wiki entity, makes the cross-lingual phrase retriever difficult to augment general NLP tasks. Moreover, unlike wiki entities, which may have fewer ambiguities, general-type phrases that are lexically identical can have different meanings depending on their contexts. Thus, accounting for polysemy (Cruse, 1986) using the context information becomes critical. Even when the lexically identical phrases share similar semantics, a more appropriate context would provide richer information for utilizing the phrases in downstream tasks (Min et al., 2019; Cao et al., 2024).

Therefore, we advocate for a new task formulation, i.e., *cross-lingual contextualized phrase retrieval*, which aims to find the cross-lingual phrase that is mostly relevant to the provided (general-type) source phrases, considering their meanings and surrounding contexts. However, achieving this goal is non-trivial, due to the scarcity of specific training data and models. Since annotating cross-lingual contextualized phrase pairs of general type is very difficult and expensive, we first introduce a data collection method, which leverages the automatically induced word alignment information from parallel sentences to extract suitable cross-lingual phrase pairs for training. This ensures that the phrases are of general type and both cross-lingual phrases and contexts are well aligned. Thereafter, we propose a **C**ross-lingual

[∗] Work done during Huayang's internship at Tencent AI Lab. Correspondence to Deng Cai and Lemao Liu.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 6562–6576
November 12-16, 2024 ©2024 Association for Computational Linguistics

Contextualized Phrase Retriever (CCPR) which is trained on the constructed dataset. The CCPR mainly employs phrase-level contrastive learning to draw cross-lingual phrases with similar contexts and meanings closer in the hidden space. At inference time, we can leverage the well-trained CCPR to build a phrase-level index and use the query phases for search.

Unlike previous works (Zheng et al., 2022), which focus solely on the task of cross-lingual phrase retrieval for wiki entities, our work also explores the potential of leveraging CCPR to augment downstream cross-lingual tasks, e.g., machine translation (MT). One critical question in front of us is how to select meaningful phrases for indexing and searching at inference time. To address this problem, we propose to learn a phrase segmentation module, which can be used to predict meaningful phrases from sentence- or passage-level retrieval data at inference time. The developed phrase segmentation module is important for ensuring the train-test consistency of CCPR.

Experiments on both of the cross-lingual contextualized phrase retrieval and a downstream task, i.e., MT, show the effectiveness of CCPR. We first evaluate CCPR on the cross-lingual contextualized phrase retrieval task, since a higher accuracy is more beneficial for downstream tasks, such as MT, cross-lingual dictionary induction, etc (Zhang and Zong, 2016; Søgaard et al., 2018). For this task, we use the human annotated cross-lingual phrase pairs as the golden truth and evaluate whether CCPR and other baselines can retrieve those golden phrases from a large-scale index. Experiments show that CCPR outperforms the baselines for at least 13 points in terms of the top-1 accuracy. We also conduct evaluation on the MT task using a large language model (LLM), e.g., LLama-2 (Touvron et al., 2023). This task aims to evaluate whether the information retrieved by CCPR can enhance LLM's ability of cross-lingual generation. Following the fashion of retrieval-augmented generation (RAG) (Lewis et al., 2020b), we simply integrate the retrieved phrase information to the input of the LLM, and compare our method with other baselines. For both X⇒En and En⇒X translation directions on the WMT16 dataset, where X is from six languages, our CCPR achieves averaged gians of 0.7 and 1.5 BERTScore points (Zhang et al., 2019), respectively. Results on other evaluation metrics, e.g., COMET (Rei et al., 2022), are consistent with BERTScore.

In summary, our contributions are three-fold:

- We propose a new formulation of dense retrieval, i.e., *the cross-lingual contextualized phrase retrieval*, which has substantial potential in augmenting cross-lingual tasks.
- We propose a **C**ross-lingual **C**ontextualized **P**hrase **R**etriever (CCPR), which uses the constructed training data, i.e., cross-lingual phrases extracted from automatically induced word alignment information, to learn both the phrase alignment and segmentation modules.
- We conduct extensive experiments on cross-lingual contextualized phrase retrieval and MT. Our method outperforms the baselines by a large margin.

2 Related Work

Dense Retrieval In text generation, dense retrieval has been widely studied at both sentence and phrase levels. Particularly after the emergence of large language models (LLMs), how to retrieve related sentences (or passages) (Izacard et al., 2021; Ni et al., 2022) for input to LLMs has became a popular research topic, namely, retrieval-augmented generation (RAG) (Lewis et al., 2020b; Karpukhin et al., 2020; Guu et al., 2020; Borgeaud et al., 2022; Asai et al., 2024). However, compared to the sentence-level retrieval, phrase-level retrieval has been shown to not only enhance retrieval accuracy but also markedly boost performance in downstream tasks (Lee et al., 2021a,b). By indexing and integrating retrieved phrases into the output of language models, Lan et al. (2023) and Cao et al. (2024) make the text generation process more attributable and accurate. These studies underline the effectiveness of phrase-level dense retrieval. Our work notably diverges from those works, because we focus more on cross-lingual tasks.

Retrieval in Cross-lingual Tasks In cross-lingual tasks, the majority of research has concentrated on sentence-level retrieval. Many research works in retrieval-augmented MT (Zhang et al., 2018; Gu et al., 2018; Xia et al., 2019; He et al., 2021) have explored retrieving bilingual sentences based on similarity of source sentences, guiding MT models to utilize retrieved target information. A notable limitation is their reliance on bilingual data for constructing the index. This requirement substantially limits the scale and diversity of data available for retrieval. In contrast, Cai et al. (2021) propose to directly retrieve relevant target sentences using source sentences to assist MT, facilitating cross-lingual retrieval using a monolingual index. Unlike the task-specific model for MT in Cai et al. (2021), many studies in dense retrieval aim to enhance the cross-lingual sentence retrieval on more languages and at a larger scale (Chidambaram et al., 2019; Conneau et al., 2020; Reimers and Gurevych, 2020; Heffernan et al., 2022; Feng et al., 2022; Cai et al., 2022).

However, cross-lingual phrase retrieval has not been extensively investigated. Bapna and Firat (2019) utilize the source side of the bilingual data for building a retrieval index and employ the similarity between source phrases for retrieval. Therefore, their indexing also suffers from the limitation of reliance on bilingual data. In addition, it uses n-grams as phrases by default, leading to plenty of meaningless and noisy spans in the index. Zheng et al. (2022) limit their cross-lingual phrase retrieval task to a specific phrase type, i.e., the wiki entity. The setting in their work overlooks the nuanced meanings that lexically identical phrases may convey in different contexts, making the retriever difficult to augment general NLP tasks. In contrast, our research aims to develop a cross-lingual contextualized phrase retriever for general-type phrases.

Additionally, there is research focused on employing nearest neighbor retrieval to support MT (Khandelwal et al., 2021; Zheng et al., 2021; Meng et al., 2022; Deguchi et al., 2023). These efforts differ from the aforementioned studies as they aim to estimate prediction distributions through nearest neighbor retrieval at each translation step.

3 Task Formulation

This work aims a cross-lingual phrase retriever that can augment downstream NLP tasks. To ensure its effectiveness, the retriever must adhere to two essential criteria. First, it should ensure that phrases in the index are of general type. Second, it must resolve the polysemy of general-type phrases by considering their contextual differences.

Thus, we introduce a new task formulation, *cross-lingual contextualized phrase retrieval*. We are given a large collection of N general-type phrases $\mathcal{P}_{index} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N\}$ in language L_y, and a general-type query phrase $\boldsymbol{q}$ in language L_x. Each phrase, by default, is associated with the context information $(\boldsymbol{c}, s, e)$, where $\boldsymbol{c}$ represents the original sentence containing this phrase, and s and e are its start and end positions in $\boldsymbol{c}$. The primary objective is to identify a cross-lingual phrase $\boldsymbol{p} \in \mathcal{P}_{index}$ that is relevant to $\boldsymbol{q}$, considering their contexts and meanings. Our formulation notably diverges from the one proposed by Zheng et al. (2022), which focuses on a restricted type of phrases, i.e., wiki entities, ignoring the nuanced meanings that general-type phrases may convey in different contexts.

This new task formulation also introduces several challenges, with the primary issue being the lack of training data. Specifically, there is a scarcity of cross-lingual phrase pairs of general type and accompanied by contextual information. Such data are crucial for training the model to recognize cross-lingual phrases that have similar contexts and meanings, allowing it to align these phrases closely in a hidden space. In addition, how to extract meaningful general-type phrases from sentences is also an open question, which is critical for employing the retriever to downstream tasks.

4 Training Data Collection

Annotating cross-lingual contextualized phrase pairs of general type is very difficult and expensive, posing a significant obstacle to training effective cross-lingual contextualized phrase retriever. However, the sentence-level parallel data in general domains are more readily available (Ng et al., 2019). Additionally, as the lower unit of phrase alignment, the word alignment has been extensive studied (Brown et al., 1993; Och and Ney, 2003; Dyer et al., 2013; Li et al., 2019; Jalili Sabet et al., 2020; Dou and Neubig, 2021; Wu et al., 2023). Therefore, to overcome the data scarcity, we propose to use a word alignment model, such as GIZA++ (Och and Ney, 2003) and neural aligner (Dou and Neubig, 2021), in order to automatically induce word alignments from parallel sentences and subsequently extract cross-lingual phrase pairs. This method allows us to produce phrase pairs that are not only of general type but also accompanied with contexts, aligning with our task requirements.

Formally, given a pair of parallel sentences $\boldsymbol{x} = \{x_1, x_2, ..., x_{|\boldsymbol{x}|}\}$, and $\boldsymbol{y} = \{y_1, y_2, ..., y_{|\boldsymbol{y}|}\}$, where $|\boldsymbol{x}|$ and $|\boldsymbol{y}|$ are the number of words in the sequences, we use a word alignment model to obtain the word alignment information of $\boldsymbol{x}$ and $\boldsymbol{y}$ (Koehn et al., 2003; Och et al., 1999). If every word in a consecutive span $\boldsymbol{x}_{i:j}$ can be aligned to a consecutive span $\boldsymbol{y}_{u:v}$, we will use them as a pair of aligned phrases $\boldsymbol{p}^x = \boldsymbol{x}_{i:j}$ and $\boldsymbol{p}^y = \boldsymbol{y}_{u:v}$, where

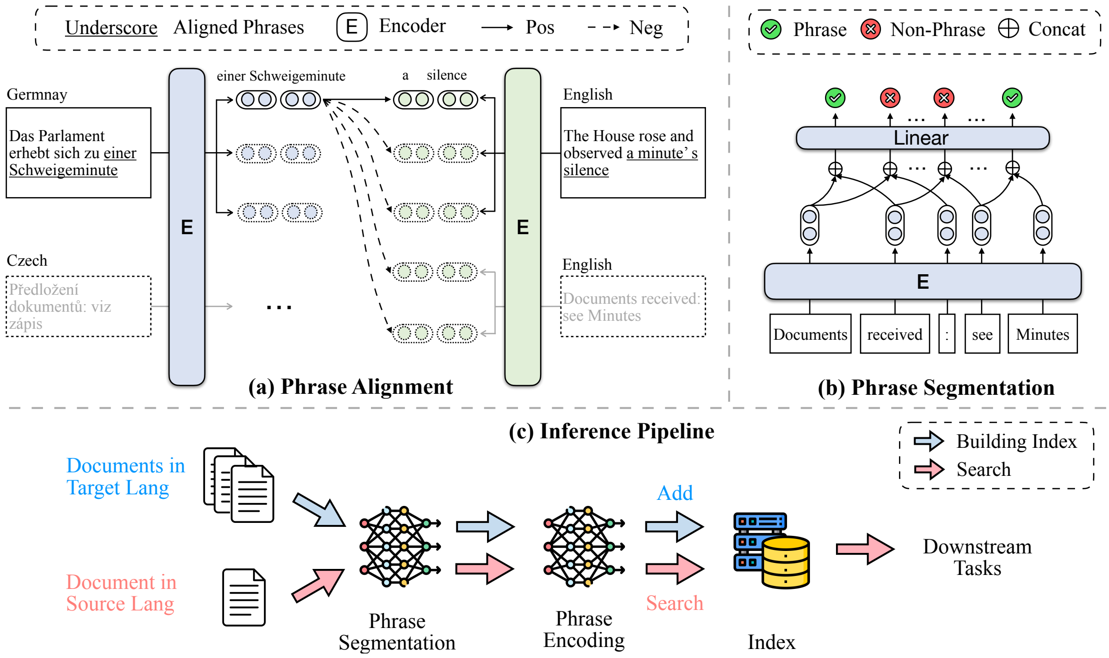

Figure 1: (a) Cross-lingual phrase alignment takes the cross-lingual phrase pairs within similar contexts as positives and other in-batch phrases as negatives. Encoders with different colors employ different dropout mask $\boldsymbol{z}$, following Gao et al. (2021). The inputs of the dual encoders are parallel sentences. (b) A linear phrase-segmentation head is used to predict whether a span is a phrase or not. (c) The pipeline of using our learned model for downstream tasks.

$1 \leq i \leq j \leq |\boldsymbol{x}|$ and $1 \leq u \leq v \leq |\boldsymbol{y}|$. Notably, phrases that meet the constraints may range from a single word to the whole sequence, and extracted phrases are allowed to have overlaps. In addition, the context information of $\boldsymbol{p}^x$ and $\boldsymbol{p}^y$ are $(\boldsymbol{x}, i, j)$ and $(\boldsymbol{y}, u, v)$, respectively. Unlike the phrase table extracted in SMT (Koehn et al., 2003; Chiang, 2005), which is static and context-independent, here the phrase mapping between $\boldsymbol{p}^x$ and $\boldsymbol{p}^y$ is unique considering their context.

As shown in Fig. 1 (a), The Germany phrase (“einer”, “Schweigeminute”) is aligned with the the English phrase (“a”, “minute”, “’s”, “silence”) within similar contexts, i.e., parallel sentences. More details about the extraction of cross-lingual contextualized phrase pairs are in App. A.1.

5 Methodology

5.1 Model Architecture

We introduce a **C**ross-lingual **C**ontextualized **P**hrase **R**etriever (CCPR) for our new formulated task. The main target of CCPR is to make the representations of cross-lingual phrases with similar contexts and meanings to be close in the hidden space. To this end, we propose a *cross-lingual phrase alignment* module based on contrastive learning, utilizing the data collected in Sec. 4 for training. In addition, one remaining problem is how to select phrases for indexing at inference time. To address this, we propose a *phrase segmentation* module that uses phrase representations to predict meaningful phrases from sentences or paragraphs. At inference time, we can leverage the learned phrase segmentation module to select phrases for indexing, ensuring the train-test consistency.

Phrase Encoding Our phrase encoder is based on a context encoder and an MLP layer. The encoding of a phrase $\boldsymbol{p} = \boldsymbol{x}_{s:e}$, is defined as:

$$\mathbf{H}^z = \mathrm{ContextEncoder}(\boldsymbol{x}, \boldsymbol{z})$$
$$\boldsymbol{h}_p^z = \mathrm{MLP}_{align}([\mathbf{H}_s^z; \mathbf{H}_e^z]) \quad (1)$$

where $\mathrm{ContextEncoder}(\boldsymbol{x}, \boldsymbol{z})$ is a Transformer model, e.g., BERT (Devlin et al., 2019; Feng et al., 2022) and RoBERTa (Conneau et al., 2020), that encodes the context $\boldsymbol{x}$ to a matrix $\mathbf{H}^z \in \mathbb{R}^{N \times d}$ with a dropout mask $\boldsymbol{z}$. Inspired by preview works (Lan et al., 2023; Lee et al., 2021a; Seo et al., 2018), we use the concatenation of $\mathbf{H}_s^z$ and $\mathbf{H}_e^z$, i.e., the start and end token of the phrase, to represent phrase $\boldsymbol{x}_{s:e}$. $\mathrm{MLP}_{align}(\cdot)$ is a function that maps the concatenation of two hidden states $\mathbf{H}_s^z$ and $\mathbf{H}_e^z$ from $\mathbb{R}^{2d}$ to $\mathbb{R}^o$, where o is the output hidden size. The encoding for phrases in context $\boldsymbol{y}$ is similar.

Cross-lingual Phrase Alignment For a batch of parallel sentences, we can extract a collection of cross-lingual phrase pairs $\mathcal{P}_{pair} = \{(\boldsymbol{p}_i^x, \boldsymbol{p}_i^y)\}_{i=1}^K$, where $\boldsymbol{p}_i^x$ and $\boldsymbol{p}_i^y$ are the positive examples of each other. Since each positive phrase pair is from parallel sentences that share similar semantics, directly learning on them may cause the model to learn a trivial shortcut. Therefore, inspired by SimCSE (Gao et al., 2021), we apply two independently sampled dropout mask $\boldsymbol{z}$ and $\boldsymbol{z}'$, to encode phrases $\boldsymbol{p}_i^x$ and $\boldsymbol{p}_i^y$, respectively, following Eq. 1. The training objective for the $x \rightarrow y$ direction is:

$$\mathcal{L}_{x\rightarrow y} = -\frac{1}{K}\sum_{i=1}^{K} \log \frac{\exp(\boldsymbol{h}_{p_i^x}^{z} \cdot \boldsymbol{h}_{p_i^y}^{z'})}{Z_x(i)}, \quad (2)$$

$$Z_x(i) = \sum_{(p_j^x, p_j^y)\in\mathcal{P}_{pair}} \exp(\boldsymbol{h}_{p_i^x}^{z} \cdot \boldsymbol{h}_{p_j^y}^{z}) \quad (3)$$

where we use all the non-paired phrases in $\mathcal{P}_{pair}$ as in-batch negatives in Eq. 3. For bidirectional symmetry, our final loss is:

$$\mathcal{L}_{align} = \mathcal{L}_{x\rightarrow y} + \mathcal{L}_{y\rightarrow x}. \quad (4)$$

The illustration of the contrastive learning is shown in Fig. 1 (a). In practice, we may use parallel data from multiple languages during training, i.e., the source contexts $\boldsymbol{x}$ in a batch may come from multiple languages, e.g., Germany, Czech, etc, which is inspired by the success of multi-lingual training in Liu et al. (2020).

Phrase Segmentation This module aims to learn how to select phrases from sentences or passages, since most retrieval data at inference time comes in sentences or passages, rather than phrases. To ensure the train-test consistency, we take the phrases extracted in Sec. 4 as positive data, and all the other spans in the corresponding contexts as negative data. Our phrase segmentation module is a binary classifier defined as follows:

$$P(T=1) = \sigma\Big(\mathrm{MLP}_{seg}\big([\mathbf{H}_i^z; \mathbf{H}_j^z]\big)\Big) \quad (5)$$

where T is the label for the span $\boldsymbol{x}_{i:j}$, where $1 \le i \le j \le |\boldsymbol{x}|$, $\sigma(\cdot)$ is an activation function, MLP_{seg} is linear layer that maps the hidden representation from $\mathbb{R}^{2d}$ to $\mathbb{R}$, and the definitions for $\mathbf{H}_i^z$ and $\mathbf{H}_j^z$ are the same as in Eq. 1. We use the label $T = 1$ for phrases and $T = 0$ for non-phrase spans. The phrase segmentation for sequence $\boldsymbol{y}$ is similar to Eq. 5. In practice, the number of non-phrase spans is significantly more than the number of phrases. To mitigate data imbalance, we employ a strategy of randomly sampling an equal number of non-phrase spans and phrases within a sentence during training (Li et al., 2020). The illustration of phrase segmentation is in Fig. 1 (b).

The training loss for the phrase segmentation is

$$\mathcal{L}_{seg} = -\frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{p}\in\mathcal{S}} \Big(T_p \log P(T_p = 1) + (1 - T_p)\log\big(1 - P(T_p = 1)\big)\Big) \quad (6)$$

where $\mathcal{S}$ is a set of spans, which are extracted from sequence $\boldsymbol{x}$ or $\boldsymbol{y}$. The learning objective of the whole model becomes:

$$\mathcal{L} = \mathcal{L}_{align} + \beta\mathcal{L}_{seg} \quad (7)$$

where β is a hyper-parameter. It is worth noting that the parameters of $\mathrm{ContextEncoder}$ used for phrase alignment and segmentation are shared.

In our preliminary studies, we investigated various phrase segmentation strategies, including n-gram segmentation and learning a Byte Pair Encoding (BPE) model (Sennrich et al., 2016; Kudo, 2018) across word boundaries. However, these methods struggle to identify meaningful phrases, leading to significant discrepancies between training and testing phases and performance declines. In Sec. 6.3, we evaluate the effect of the learned phrase segmentation on MT task.

5.2 Inference Pipeline

The pipeline of employing our cross-lingual contextualized phrase retriever for downstream NLP tasks involves two main steps: index building and searching, as shown in Fig. 1 (c).

A significant benefit of our retriever is its ability to leverage monolingual data in the target language, eliminating the need for bilingual data to construct the index at inference time (Zhang et al., 2018; Gu et al., 2018; Bapna and Firat, 2019). Given the retrieval data in target language, our retriever begins by segmenting sentences (or passages) into phrases. When segmenting a sentence (or passage), we will enumerate all the possible spans within it and calculate their probabilities of being meaningful phrases, according to Eq. 5. We select all the spans whose probabilities are larger than a threshold for indexing. It is worth noting that overlapping phrases are allowed at the phrase segmentation time. The selected phrases are encoded into hidden representations according to Eq. 1, which will be used to construct a dense-retrieval index.

... Explanations about the phrase information ...

Germany Phrase: Premierminister
Potential Translation: prime ministers
Context: ... leaders stayed away, including the [[prime ministers]] of Canada and India ...

Germany Phrase: Indiens und Japans
Potential Translation: India and Japan
Context: ... right governments in [[India and Japan]], the weakening clout of Arab oil ...

Germany Phrase: trafen
Potential Translation: met
Context: ... Obama and Abe [[met]] with Japanese university students ...

Germany Phrase: Tokio
Potential Translation: Tokyo
Context: ... Canadian officials privately point fingers at [[Tokyo]] ...
...

Based on the provided information of phrase translation, please faithfully translate the following sentence from Germany into English:

Germany: Die Premierminister Indiens und Japans trafen sich in Tokio.

English: India and Japan prime ministers meet in Tokyo

Figure 2: An example instruction for a large-language-model based translator augmented by our method. The segmented phrases in the source sentence are in blue text. The retrieved translation of the phrases are in green, and their appearances in the contexts are marked by "[[]]". The text after "English:" is the reference, which is for illustration and will not appear in the instruction.

For querying, we first process the query sentence (or paragraph) in a source language similarly, i.e., segmenting and encoding. Subsequently, we use the encoded query phrases to directly search for the related cross-lingual phrases using Maximum Inner Product Search (MIPS). Because of the advanced data structure and search algorithm (Malkov and Yashunin, 2018; Douze et al., 2024; Johnson, 2022), the search step is highly efficient (more details in App. A.3). The retrieved information, including the target phrase, its accompanied context, and its positions in the context, can then be integrated to downstream tasks to improve the performance. As shown in Fig. 2, the retrieved phrases along with their surrounding contexts are integrated into the instructions given to an LLM for tasks such as MT.

Model	De⇒En	Ro⇒En	Cs⇒En	AVG.
XLMR	1.5	0.0	4.0	1.8
mBERT	8.0	3.0	7.5	6.1
MUSE	34.5	30.0	45.5	36.6
LaBSE	52.0	33.5	61.0	48.8
CPR-XLMR	54.0	63.0	73.0	63.3
CPR-LaBSE	57.0	60.0	78.5	65.1
CCPR-XLMR	74.5	**73.5**	82.5	76.8
CCPR-LaBSE	**75.0**	72.5	**88.0**	**78.5**

Table 1: Cross-lingual contextualized phrase retrieval. We use the accuracy@1 as our evaluation metric. The CCPR and CPR denote contextualized and context-independent cross-lingual phrase retriever, respectively. The Best results are highlighted in **bold text**.

6 Experiments

We propose to evaluate our model on cross-lingual contextualized phrase retrieval and MT.

6.1 Implementation Details

Training Data Collection We use all the bilingual training datasets of WMT16 on Huggingface[1] to train our **C**ross-lingual **C**ontextualized **P**hrase **R**etriever (CCPR). The WMT16 dataset has six language pairs, inclduing De-En, Cs-En, Fi-En, Ru-En, and Tr-En. Because of the efficiency and the satisfactory performance of GIZA++ (Och and Ney, 2003; Dou and Neubig, 2021), we first use the GIZA++ software[2], i.e., IBM-4 model (Brown et al., 1993), to induce the word alignment for each pair of parallel sentences, and then extract the cross-lingual phrase pairs. More details about phrase extraction are in App. A.1. The final training dataset contains 1.3 billion cross-lingual phrase pairs extracted from 10 million parallel sentences.

Model In our work, we train two variants of our model, CCPR-XLMR and CCPR-LaBSE, whose parameters are initialized from XLM-RoBERTa-base (XLMR) (Conneau et al., 2020) and LaBSE (Feng et al., 2022), respectively. The hidden size d of both models is 768. The output o of $\text{MLP}_{align} : \mathbf{R}^{2d} \rightarrow \mathbf{R}^{o}$ in Eq. (1) is 128, which plays a critical role in reducing the memory cost of our phrase index while maintaining comparable performance. At inference time, we use the FAISS library[3] (Douze et al., 2024) to build our retrieval index. We use the `FlatIP` as our index type. Training details are App. in A.2

[1] https://huggingface.co/datasets/wmt16
[2] https://github.com/moses-smt/mgiza
[3] https://github.com/facebookresearch/faiss

Model (X⇒En)	De⇒En	Cs⇒En	Fi⇒En	Ru⇒En	Ro⇒En	Tr⇒En	AVG.
LLAMA-2-7B	75.42	69.03	66.13	71.65	72.89	55.10	68.37
+ XLMR	75.32	69.15	66.06	71.74	73.07	55.49	68.47
+ LABSE	75.74	69.60	66.81	72.12	73.22	56.76	69.04
+ CCPR-XLMR	75.84	**70.11**	66.70	71.49	73.53	58.44	69.35
+ CCPR-LABSE	**75.96**	70.05	**67.17**	**72.28**	**73.57**	**59.22**	**69.70**
Model (En⇒X)	**En⇒De**	**En⇒Cs**	**En⇒Fi**	**En⇒Ru**	**En⇒Ro**	**En⇒Tr**	**AVG.**
LLAMA-2-7B	63.42	52.32	48.96	84.18	82.99	58.47	65.05
+ XLMR	64.03	52.34	49.04	84.08	82.99	58.66	65.19
+ LABSE	64.26	52.90	49.76	84.28	83.05	60.02	65.71
+ CCPR-XLMR	**64.42**	54.53	51.12	84.93	83.18	63.40	66.92
+ CCPR-LABSE	64.29	**54.93**	**51.54**	**84.97**	**83.36**	**64.07**	**67.19**

Table 2: MT on the test sets of WMT16. For all the retrieval-based method, the index is built on the monolingual newscrawl data of the target language. We use the BERTScore (Zhang et al., 2019) as the evaluation metric. The best performance of each translation direction is highlighted in **bold text**. Results on additional evaluation metrics, e.g., COMET (Rei et al., 2020), are in App. C.4

6.2 Cross-lingual Contextualized Phrase Retrieval

As outlined in Sec. 3, given a query phrase and its context in source language, the objective of this task is to identify the most relevant cross-lingual phrase from a large-scale index in target language, considering the context information.

Setup We build the test set based on the human annotated word alignment data (Jalili Sabet et al., 2020). We process the word alignment data on three language pairs, De⇒En (Ghader and Monz, 2017), Ro⇒En (Mihalcea and Pedersen, 2003), and Cs⇒En (Mareček, 2011). For the Cs⇒En dataset, we only leverage the data under the "pcedt" split, which are data from The Wall Street Journal (WSJ). For each language pair, we instruct human annotators to identify 200 high-quality phrase pairs. The details of the human annotation process are discussed in App. B.1. These annotated source phrases serve as queries, while their aligned target phrases help construct the index. To mimic real-world conditions, where an index contains extra data, we use 9.6 million English phrases sampled from the WMT16 training dataset as the extra data.

We compare our models, CCPR-XLMR and CCPR-LABSE, with several baselines. The first two baselines XLMR (Conneau et al., 2020) & MBERT (Devlin et al., 2019) [4] are models trained on multilingual data. We use them to encode the context and represent the phrases by concatenating the hidden representations of their start and end tokens. The second two baselines MUSE (Chidambaram et al., 2019) & LABSE (Feng et al., 2022) are trained on cross-lingual data, and we use the model weights released by SBERT[5] (Reimers and Gurevych, 2020) to encode the contexts. To evaluate the effect of context-awareness, we propose two context-independent counterparts, i.e., CPR-XLMR and CPR-LABSE. Compared with our contextualized models, the only difference is that the CPR-X methods use cross-lingual phrase pairs in semantically different contexts as positive examples, similar to the setup in Zheng et al. (2022). We build an independent index for each model. In this task, we use the top-1 accuracy to evaluate our methods and baselines.

Model	Index	En⇒Ro	En⇒Cs	En⇒Tr	AVG.
CCPR-XLMR	Train	82.89	53.66	62.03	66.19
CCPR-XLMR	NC	**83.18**	**54.53**	**63.40**	**67.03**
CCPR-LABSE	Train	83.19	54.23	63.22	66.88
CCPR-LABSE	NC	**83.36**	**54.93**	**64.07**	**67.45**

Table 3: Analysis about the retrieval data for indexing. The "Train" and "NC" indicate using the target sides of WMT16 training data and monolingual newscrawl data to build the index, respectively. We use BERTScore as our evaluation metric.

Results Tab. 1 shows that our models significantly outperform baselines in cross-lingual contextualized phrase retrieval, highlighting the capability of our models to accurately identify relevant cross-lingual phrases while being context-aware. In contrast, the CPR-LABSE and CPR-XLMR, i.e., the context-independent baselines inspired by Zheng et al. (2022), perform notably worse.

[4] https://huggingface.co/models

[5] https://www.sbert.net

6.3 Machine Translation

The evaluation on MT task is to assess if the information retrieved by our models can enhance performance on downstream tasks. Rather than training a new MT model (Vaswani et al., 2017), we propose to integrate the information to the prompt of the LLM, e.g., Llama-2 (Touvron et al., 2023), following the fashion of retrieval-augmented generation (RAG) (Lewis et al., 2020a). Thus, we can fairly evaluate the effect of the retrieved information.

Setup We choose LLAMA-2-7B (Touvron et al., 2023) model as the backbone model for MT. We use two of our phrase retrieval models for this task, i.e., CCPR-LABSE and CCPR-XLMR, as discussed in Sec. 6.1. We also consider a multilingual and a cross-lingual sentence retrieval models as baseline methods, i.e., the XLM-RoBERTa-base (XLMR) (Conneau et al., 2020) and LABSE (Feng et al., 2022). The two sentence-level baselines are proposed to evaluate whether the phrase-level retrieval information is more beneficial for augmenting LLM's cross-lingual ability.

We integrate the retrieval information into input of the LLM. An example of the prompt of our method is shown in Fig. 2, where the retrieved target phrases with their contexts are presented along with the source phrases in the prompt. The presented contexts are truncated to no more than 100 characters in our method. More details about prompts of our method and other baselines are shown in App. C.1. For our methods, when building the index, we first use the learned segmentation module (Sec. 5.1) to extract contextualized phrases from the retrieval data and then encode them into dense vectors using our models. For the XLMR and LABSE baselines, we directly use them to encode sentences in retrieval data and build a sentence-level index. For all methods, we use the monolingual newscrawl data of 2016 [6] as the retrieval data. More setup details are in App. C.

Main Results As shown in Tab. 2, both of our CCPR-XLMR and CCPR-LABSE outperform the baseline methods when assisting the LLM with the information from cross-lingual contextualized phrase retrieval. Especially on some low-resource directions, such as the Tr⇔En, En⇒Cs, En⇒Fi, CCPR-LABSE outperforms other baselines by more or around 2 BERTScore points. This indicates that the the contextualized cross-lingual phrase information is more critical for improving the performance on low resource languages, because the percentage of Cs, Fi, and Tr data in Llama-2 training are all less than 0.03% (Touvron et al., 2023). Consistent results on COMET (Rei et al., 2022) are shown in App. C.4.

[6] https://data.statmt.org/news-crawl/${LANG}/news.2016.${LANG}.shuffled.deduped.gz

Segmentation	En⇒Ro	En⇒Cs	En⇒Tr	AVG.
N-gram	83.16	53.42	60.74	65.77
PSM	**83.36**	**54.93**	**64.07**	**67.45**

Table 4: Analysis of phrase segmentation methods. The N-gram method uses all available 5-grams for indexing and searching, whereas PSM employs the learned phrase segmentation module. The extracted phrases are encoded by the same encoder, i.e., CCPR-LABSE. We use BERTScore as the evaluation metric.

In addition, we find a good correlation between the performance of our cross-lingual contextualized phrase retrieval task and the MT task, showing that higher retrieval accuracy leads to superior results in downstream applications.

Analysis In Tab. 2, all retrieval-based methods construct their index using the monolingual newscrawl data, which is approximately six times the size of the bilingual training data. It highlights a key benefit of our cross-lingual phrase retrieval: the ability to utilize extensive monolingual resources to build the index (Cai et al., 2021). Therefore, in Tab. 3, we evaluate the differences of indexing on monolingual data and the bilingual data. We observe that exploring the vast monolingual data leads to significantly better performance. In addition, although the N-gram is widely used for phrase segmentation in previous works (Bapna and Firat, 2019; Lee et al., 2021a,b; Lan et al., 2023), our experiments demonstrate that a learned phase segmentation is more suitable for augmenting cross-lingual tasks, as shown in Tab. 4.

7 Conclusion & Future Works

This paper introduces a novel approach to dense retrieval, i.e., cross-lingual contextualized phrase retrieval, focusing on resolving phrase polysemy by utilizing contextual information. The main challenge identified is the scarcity of training data, specifically cross-lingual phrase pairs with context. To overcome this, we use a word alignment model to derive such phrase pairs from parallel sentences. We then present the Cross-lingual Contextualized Phrase Retriever (CCPR), which employs contrastive learning to effectively capture similar meanings and contexts of cross-lingual phrases. Our extensive testing across retrieval and machine translation tasks shows that CCPR significantly outperforms existing baselines.

8 Limitations

While our cross-lingual contextualized phrase retrieval holds substantial potential, it is not without limitations. Notably, the phrase-level index required by our approach is considerably larger than that of a sentence-level index, given that each sentence can encompass numerous phrases. This expansion necessitates increased disk space for storing the index, requiring additional engineering techniques to maintain the scalability of our cross-lingual contextualized phrase retriever (CCPR), e.g., index quantization or sharding. Furthermore, to improve the performance of large language models (LLMs) on cross-lingual tasks, it becomes necessary to integrate the information from multiple retrieved phrases of the query sentence into the LLM input. This integration process can lead to a rise in the inference costs associated with LLMs.

9 Ethical Statement

We focus on leveraging cross-lingual contextualized phrase retrieval to augment the performance of downstream NLP tasks. We emphasize that our model is strictly designed and applied in a manner that avoids the generation of sensitive information, such as disinformation or content aimed at deceiving individuals. Furthermore, we assure that all data utilized for the training and evaluation of our model have been sourced from publicly accessible datasets. Our commitment to ethical research ensures our work benefits the NLP field responsibly, without compromising the ethical standards.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. 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 1921–1931, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2022. Retrofitting multilingual sentence embeddings with Abstract Meaning Representation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6456–6472, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. 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 7307–7318, Online. Association for Computational Linguistics.

Bowen Cao, Deng Cai, Leyang Cui, Xuxin Cheng, Wei Bi, Yuexian Zou, and Shuming Shi. 2024. Retrieval is accurate generation. In *The Twelfth International Conference on Learning Representations*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

D Alan Cruse. 1986. *Lexical semantics*. Cambridge university press.

Hiroyuki Deguchi, Taro Watanabe, Yusuke Matsui, Masao Utiyama, Hideki Tanaka, and Eiichiro Sumita. 2023. Subset retrieval nearest neighbor machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–189, Toronto, Canada. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2018. Search engine guided neural machine translation. In *AAAI Conference on Artificial Intelligence*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. 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 3170–3180, Online. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Christian Johnson. 2022. Binary encoded word mover's distance. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 167–172, Dublin, Ireland. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Tian Lan, Deng Cai, Yan Wang, Heyan Huang, and Xian-Ling Mao. 2023. Copy is all you need. In *The Eleventh International Conference on Learning Representations*.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of llms.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2021a. Learning dense representations of phrases at scale. 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 6634–6647, Online. Association for Computational Linguistics.

Jinhyuk Lee, Alexander Wettig, and Danqi Chen. 2021b. Phrase retrieval learns passage retrieval, too. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3661–3672, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy. Association for Computational Linguistics.

Yangming Li, Shuming Shi, et al. 2020. Empirical analysis of unlabeled entity problem in named entity recognition. In *International Conference on Learning Representations*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

David Mareček. 2011. Automatic alignment of tectogrammatical trees from czech-english parallel corpus.

Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. 2022. Fast nearest neighbor machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 555–565, Dublin, Ireland. Association for Computational Linguistics.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. A discrete hard EM approach for weakly supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2851–2864, Hong Kong, China. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Minjoon Seo, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Phrase-indexed question answering: A new challenge for scalable document comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 559–564, Brussels, Belgium. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11084–11099, Toronto, Canada. Association for Computational Linguistics.

Mengzhou Xia, Guoping Huang, Lemao Liu, and Shuming Shi. 2019. Graph based translation memory for neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7297–7304.

Jiajun Zhang and Chengqing Zong. 2016. Bridging neural machine translation and bilingual dictionaries. *arXiv preprint arXiv:1610.07272*.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Heqi Zheng, Xiao Zhang, Zewen Chi, Heyan Huang, Yan Tan, Tian Lan, Wei Wei, and Xian-Ling Mao. 2022. Cross-lingual phrase retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4193–4204, Dublin, Ireland. Association for Computational Linguistics.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. Adaptive nearest neighbor machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 368–374, Online. Association for Computational Linguistics.

... Explanations about the retrieved sentence ...

Related English Sentence: The Prime Ministers of India and Pakistan recently met in Pakistan to discuss the question.
...

Based on the provided related sentence, please faithfully translate the following sentence from Germany into English, and do not alter its meaning:

Germany: Die Premierminister Indiens und Japans trafen sich in Tokio.

English: India and Japan prime ministers meet in Tokyo

Table 6: An exemplar instruction of the LLM augmented by cross-lingual sentence retrieval. The text after "English:" is the reference, which will not appear in the instruction.

Please faithfully translate the following sentence from Germany into English, and do not alter its meaning:
Germany: Die Premierminister Indiens und Japans trafen sich in Tokio.
English: India and Japan prime ministers meet in Tokyo

Table 7: An exemplar instruction of the vanilla LLM. The text after "English:" is the reference, which is just for illustration and will not appear in the instruction.

A Implementation Details

A.1 Extraction of Cross-lingual Phrase Pairs

Because of the efficiency and the satisfactory performance of GIZA++ (Och and Ney, 2003; Dou and Neubig, 2021), we first use the GIZA++ software[7], i.e., IBM-4 model (Brown et al., 1993), to induce the word alignment for each pair of parallel sentences, and then extract the cross-lingual phrase pairs. Notably, we did not induce the phrase table as in Phrase-based Statistical Machine Translation (PBSMT) (Koehn et al., 2003). Instead, in our setting, we extract the cross-lingual consecutive spans that are aligned in a pair of parallel sentence as cross-lingual phrase pairs. Each of the cross-lingual phrase pairs is associated with the surrounding contexts. We filter out the phrases which begin or end with words whose frequency is more than 30k in the training dataset. We also filter out the phrases that contain only numbers and punctuations.

[7] https://github.com/moses-smt/mgiza

A.2 Training of CCPR

Each model is trained on the mixture of all language pairs of WMT16. We train our model on 8 V100 GPUs for 20K steps, where the learning rate is 5e-5, dropout rate is 0.2, batch size on each device is 64. The β for the phrase segmentation loss in Eq. 5 is 1. Those hyper-parameters are tuned on the validation dataset.

A.3 Inference Latency

In our experiments of cross-lingual contextualized phrase retrieval, retrieving the top-32 nearest neighbors for 1000 source phrases took only 0.04 seconds on 4 V100 GPUs, thanks to FAISS's high parallelism on GPUs (Douze et al., 2024). We believe this retrieval latency is adequate for most real-time retrieval scenarios.

B Cross-lingual Contextualized Phrase Retrieval

B.1 Human Annotation

We asked human annotators to label 200 high-quality phrase pairs as the golden truth data for each of the language pairs, i.e., De⇒En, Ro⇒En, and Cs⇒En. Hiring human annotators to annotate those bilingual data is expensive and time consuming. However, fortunately, some bilingual datasets already provide the human-annotated word alignment (Ghader and Monz, 2017; Mihalcea and Pedersen, 2003; Mareček, 2011). Therefore, in our task, we only ask the human annotators, three authors of our work, to annotate the English side of the data, i.e., whether a English span is a high-quality and meaningful phrase or not.

More concretely, we first use heuristic rules as discussed in Sec. 6.1 to collect an initial set of cross-lingual phrase pairs. For each phrase pair, we ask the human annotator to answer three questions:

1. If the phrase is a single word, whether it is informative in the context?

2. If the phrase has multiple words, whether the semantics of this phrase is complete and informative? For instance, "local authorities and large" is not a phrase with complete semantics. In addition, "Of course" has complete semantics but is not informative.

Model	De-En	Cs-En	Fi-En	Ru-En	Ro-En	Tr-En	AVG.
LLAMA-2-7B	86.019	82.509	84.955	83.616	85.005	77.551	83.27
+ XLMR	85.915	82.571	84.889	83.575	84.984	77.809	83.29
+ LABSE	86.079	82.840	85.136	83.750	85.059	78.552	83.56
+ CCPR-XLMR	**86.217**	**83.079**	85.231	83.745	85.285	79.502	83.84
+ CCPR-LABSE	86.202	83.014	**85.342**	**83.855**	**85.322**	**80.123**	**83.97**
	En-De	En-Cs	En-Fi	En-Ru	En-Ro	En-Tr	
LLAMA-2-7B	82.396	79.131	82.516	84.191	82.588	58.959	78.29
+ XLMR	82.821	78.888	82.442	84.097	82.802	59.095	78.35
+ LABSE	83.128	79.533	83.214	84.403	82.849	61.183	79.05
+ CCPR-XLMR	**83.529**	80.517	83.962	**85.377**	83.776	65.975	80.52
+ CCPR-LABSE	83.042	**80.860**	**84.095**	85.308	**84.031**	**67.060**	**80.73**

Table 5: Machine Translation. COMET.

	En⇒De	**En⇒Cs**	**En⇒Ro**
CoT	55.20	47.02	80.22
CCPR-LaBSE	64.29	54.93	83.36

Table 8: Translation results based on LLAMA-7-7B. For the CoT method, we use the template in Figure 9, and all the rest setups are the same as other methods in our work. We use the BERTScore as the metric

If the answer of any questions is true, then we add this phrase to a pool of high-quality phrases. For those phrases, overlaps are allowed. We finally randomly sampled 200 annotated phrases form the pool for each language.

C Machine Translation

C.1 Translation Instructions

The instructions for the vanilla large language model (LLM), e.g., LLAMA-2, and the LLM augmented by cross-lingual sentence retrieval are shown in Tab. 6 and 7, respectively.

C.2 Building Index

The indexing strategies diverge between the baseline methods and our approach. For sentence-level baselines, i.e., XLMR and LABSE, the indexing is straightforward, directly using the sentences. In contrast, our method employs a learned phrase segmentation module to extract phrases from sentences for indexing. When applying our model to MT, we set the phrase segmentation thresholds in Eq. 5 to 0.7 for indexing and 0.9 for querying. The rationale behind a lower indexing threshold is to populate the index with a broader array of phrases, enhancing its comprehensiveness. Conversely, a higher threshold for query documents ensures that the segmented phrases are more accurate. These hyper-parameters were tuned using the validation dataset from WMT16.

C.3 Format Alignment

Since the output format of the Llama-base model is hard to control and the Llama-chat model, i.e., the model after instruction-tuning, may reject to translate a sentence that contains negative words, we decide to fine-tune the Llama base model using LoRA (Hu et al., 2022) for 100 steps to align the translation format. We use Platypus[8] (Lee et al., 2023) as our code base. We follow all the hyper-parameters of Platypus to fine-tune the LoRA module. The data we used for fine-tuning contain 1,800 parallel sentences randomly sampled from the WMT16 training set. For each method, we only fine-tune one model for the X⇒En (or En⇒X) direction, where X is the non-English language.

C.4 More Results

More results on evaluated by COMET (Rei et al., 2020) are shown in Tab. 5, which is consistent with the results evaluated by BERTScore (Zhang et al., 2019).

Additional, to evaluate that the effectiveness of the retrieved information, we also compare with a Chain-of-Though (COT) approach (Wei et al., 2022). As shown in Figure 8, simply translating phrases before sentence translation does not achieve the expected performance gain and is slightly worse than the baseline approach. In contrast, our approach outperforms this method by a large margin. One hypothesis is that our approach

[8] https://github.com/arielnlee/Platypus

Instruction:
Please faithfully translate the following sentence from English into Germany, and do not alter its meaning.

English:
Sumptuous accommodations, the ultimate in Guest Services and a relaxing, professional atmosphere are beckoning at The Ambiance.

Output:
Let's translate the sentence step by step. Firstly, we can translate the English phrases as follows:

English Phrase: Sumptuous accommodations
Germany Phrase: prächtige Unterkünfte

(more phrases)
...
Based on the phrase-level translation, the sentence-level Germany translation is:

Kostspielige, prächtige Unterkünfte, Serviceleistungen höchsten Standards und eine entspannende Atmosphäre erwarten Sie im The Ambiance.

Table 9: An exemplar template for the translation with Chain-of-Thought. The text is the reference, which is just for illustration and will not appear in the input.

is able to leverage external knowledge, which is more helpful for the translation task.