Multimodal Cross-lingual Phrase Retrieval

Chuanqi Dong[2], **Wenjie Zhou**[2], **Xiangyu Duan**[*1,2], **Yuqi Zhang**[3] **and Min Zhang**[1,2]
[1]Institute of Aritificial Intelligence, Soochow University, China
[2]School of Computer Science and Technology, Soochow University, China
[3]Alibaba DAMO Academy
{cqdong,wjzhou223}@stu.suda.edu.cn, xiangyuduan@suda.edu.cn
chenwei.zyq@alibaba-inc.com, minzhang@suda.edu.cn

Abstract

Cross-lingual phrase retrieval aims to retrieve parallel phrases among languages. Current approaches only deals with textual modality. There lacks multimodal data resources and explorations for multimodal cross-lingual phrase retrieval (MXPR). In this paper, we create the first MXPR data resource and propose a novel approach for MXPR to explore the effectiveness of multi-modality. The MXPR data resource is built by marrying the benchmark dataset for textual cross-lingual phrase retrieval with Wikimedia Commons, which is a media store containing tremendous texts and related images. In the built resource, the phrase pairs of the textual benchmark dataset are equipped with their related images. Based on this novel data resource, we introduce a strategy to bridge the gap between different modalities by multimodal relation generation with a large multimodal pre-trained model and consistency training. Experiments on benchmarked dataset covering eight language pairs show that our MXPR approach, which deals with multimodal phrases, performs significantly better than pure textual cross-lingual phrase retrieval.We release the code and data at https://github.com/sdongchuanqi/MXPR

Keywords: Multimodal Cross-lingual Phrase Retrieval, Multimodal Relation, Consistency Training

1. Introduction

Cross-lingual phrase retrieval is a task of finding parallel phrases in a bilingual or multilingual phrase pool (Zheng et al., 2022). It is beneficial for cross-lingual applications such as named entity linking (Sil et al., 2018), question answering (Rücklé et al., 2019), and global e-commerce (Li et al., 2020b). It usually leverages sentence level encoding to extract phrase level representation, and explores the parallelism between the representations of source phrase and target phrase. Although it achieves good performance across a variety of languages, it only deals with textual modality, leaving the question of the effectiveness of using multimodal information unanswered.

Multimodal information have been studied in interdisciplinary directions such as multimodal machine translation (Yao and Wan, 2020), multimodal sentiment analysis (Soleymani et al., 2017), and multimodal named entity recognition (Moon et al., 2018). Images or speeches used in these researches are beneficial to improve the performance, proving the potential of using multimodal information.

In the area of cross-lingual phrase retrieval, however, there are neither multimodal data resources nor the corresponding approaches for multimodal cross-lingual phrase retrieval (MXPR). To solve this issue, we create the first MXPR data resource and propose a novel framework for MXPR. The data resource is based on the textual benchmark dataset WikiXPR [1] extracted from Wikipedia. We equip each phrase pair in WikiXPR with the related image by using an image retrieval engine in Wikimedia Commons [2], which is a large-scale media file repository. Since both phrase pairs and images are from Wikimedia, it ensures almost all phrase pairs have their related images.

Given the created MXPR data resource, our task faces a challenge to conduct cross-lingual phrase retrieval with image information due to their different modalities. Recently, along with the development of multimodal large language model(M-LLM), significant progresses have been made in various cross-modal tasks such as visual question answering (Li et al., 2023), image-text retrieval (Ye et al., 2023). Inspired by the success of M-LLM, we introduce it into our framework by instructing it to generate relations between images and phrases, so that the two modalities are aligned better. Considering that the generated contents by M-LLM may contain noises, we propose a consistency training scheme that balances between the textual and image modalities to alleviate the noise problem.

Experiments on three categories of benchmarked cross-lingual phrase retrieval tasks, i.e., bilin-

*Corresponding author

[1]https://github.com/cwszz/XPR

[2]https://commons.wikimedia.org/wiki/Main_Page

LREC-COLING 2024, pages 11917–11927

Table 1: Number of phrase pairs equipped with the related images in MXPR data resource.

	En-Zh	En-Fr	En-Ar	En-De	En-Es	En-Ja	En-Ko	En-Ru
training set	8247	1315	4177	1914	1327	14545	2114	5188
validation set	2751	433	1387	638	442	4849	708	1730
test set	2756	435	1391	637	441	4853	707	1726

gual retrieval, multilingual retrieval, and zero-shot transfer on all eight language pairs, show that our MXPR approach performs significantly better than the approaches using only textual modality, which manifests the effectiveness of the multimodal information. Our instructed multimodal relation and consistency training consistently improve the performance. If phrase pairs are coupled with random images, the performance decreases significantly, which indicates that image information is truly important for improving the performance. In summary, our contributions are:

- We create the first MXPR data resource.
- Instructed multimodal relation and consistency training are proposed to well align image and phrase modalities.
- Comprehensive experiments demonstrate the effectiveness of our created data resource and the proposed multimodal approach, which lead to the significant improvement over the approaches using only textual modality.

2. MXPR Data Resource Creation

Although no dataset of multimodal cross-lingual phrase pairs exist, there are image retrieval engine and unimodal cross-lingual phrase pairs that can be utilized together to build MXPR dataset. So, we combine an image retrieval engine from Wikimedia Commons and the textual cross-lingual phrase pair dataset WikiXPR (Zheng et al., 2022) to build our MXPR data. The same origin of Wikipedia ensures them relate to each other well.

In particular, Wikimedia Commons is a large-scale repository storing freely licensed educational media content including images, sound, and video clips. It provides an image retrieval engine that can retrieve related images in Wikimedia Commons for a phrase input. WikiXPR is an English-centric dataset consisting of phrase pairs for eight language pairs. It also contains example sentences for each monolingual phrase for enhancing the phrase representation.

For each English phrase in a phrase pair of WikiXPR, we retrieve its related image in Wikimedia Commons. In the end, 64,711 phrase pairs out of a total of 65,400 phrase pairs in WikiXPR can find the related images. The covering rate is 98.94%.

Table 2: Statistics of the three categories of the relations between phrases and images on the test sets.

	equivalent	*related*	*unrelated*
En-De	126	494	17
En-Fr	85	343	7
En-Es	84	345	12
En-Ar	362	1004	25
En-Ko	129	565	13
En-Ru	413	1290	23

The training, validation, and test sets are split as same as WikiXPR. Statistics are listed in Table 1.

Investigation on Multimodal Relation In the data resource, not all phrases and their images are equivalent to each other. For example, the phrase 'Japanese Parliament' is somehow an abstract entity, which does not have concrete image. The retrieved image may just have a relation to the phrase, but is not fully equivalent to the phrase. So, we investigate the relations between the phrases and images through manual labeling on the test sets. We divide the relation into three categories: *equivalent*, *related*, and *unrelated*, and constitute the labeling rule for each category. Basically, if a phrase is grounded to the whole or major part of its retrieved image, then the phrase and the image are labeled *equivalent*. If a phrase is grounded to a small part of its retrieved image or the retrieved image is corresponding to a part of the phrase meaning, then they are labeled *related*. If a phrase can not be grounded to any part of its retrieved image, then they are labeled *unrelated*.

Four annotators participated in this manual labeling process. The experienced annotator writes the guideline, then the other three annotators try to annotate a small portion according to the guideline and send to the experienced annotator to check the quality. In the last, the three annotators begin the formal annotation, and the experienced annotator checks the quality through sampling. The process iterates until no problem can be found in the sampling. Figure 2 presents examples of the three relation categories.

Table 2 lists the statistics of the three categories of the multimodal relation on the test sets. It shows that the majority of the multimodal relations is *re-*

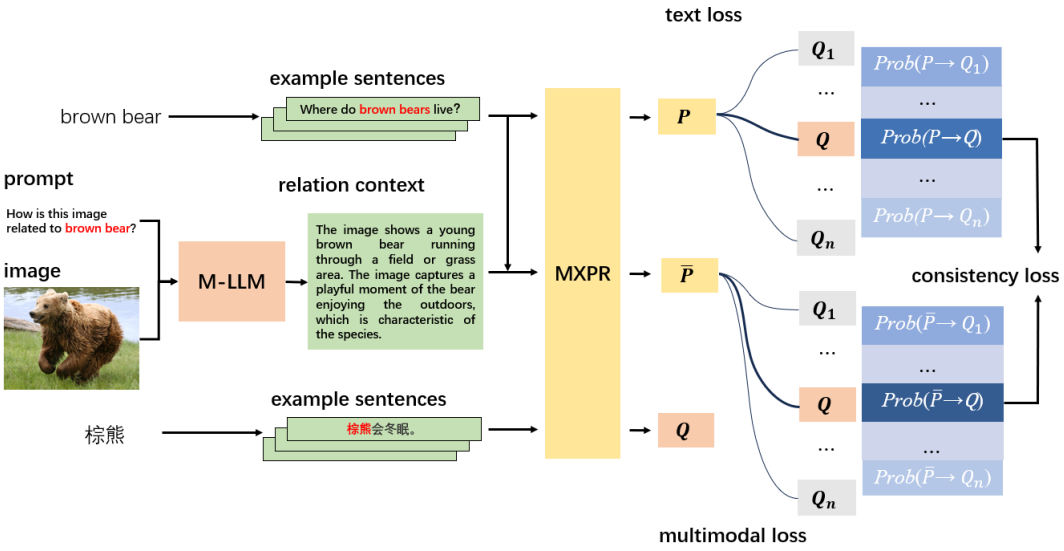

Figure 1: MXPR Framework. The phrase and its related image are used to prompt an M-LLM to generate relation text at first, then MXPR encoder is trained to align source phrase and target phrase representations given the relation text and example sentences as input.

lated, while *equivalent* is significantly less than *related*, and *unrelated* only takes up a tiny portion. Such relation distribution indicates that the best way of using multimodal information in MXPR data resource is utilizing the multimodal relation instead of only using images equivalent to phrases. In fact, we have tried using image caption systems for describing the image, but we found that the captions often missed the phrases since many images are just related to the phrases, resulting in that there is no guarantee of the appearances of the phrases in the captions. So, we resort to focus on the relations between the images and the phrases. We analyze the effects of the three categories of the relation in experimental section 5.1, which shows that the relations of *equivalent* and *related* contribute to the performance improvement, while the relation of *unrelated* harms the performance or is trivial to the task.

3. Approach

Given the above MXPR data resource, we propose to train an MXPR network that can retrieve translation of a new source phrase with the help of the corresponding image information.

3.1. MXPR Framework

MXPR framework builds a common multimodal representation space for both source side and target side phrases so that phrases in a pair are closest neighbors in the space. The challenge is that the textual phrases and their related images are in different modality, resulting in the difficulty in building the common space. Previous researches (Caglayan et al., 2019; Yao and Wan, 2020; Moon et al., 2018) built the common space by cross-modal encoding of sentences or long texts with images, but they are not fit for our scenario consisting of phrases and images because phrases are usually too short for sufficient textual encoding. MXPR solves this difficulty through instructed multimodal relation and consistency training. The solution is illustrated in Figure 1.

Take the phrase 'brown bear' and its image for example. In the first step as shown in the left part of Figure 1, MXPR framework converts the image information into textual information by instructing an M-LLM to generate the multimodal relation. The benefits are twofold: 1) The phrases are transformed into long texts which carry image information. The long texts are easier to be encoded with rich contexts than the short phrases. 2) The challenge of cross-modal encoding is bypassed that only textual modality is left after conversion. In the second step as shown in the right part of Figure 1, both multimodal relation texts and example sentences are fed into MXPR encoder to get representations of phrases on both source and target side. The encoder is trained to optimize the alignment between the source and the target phrases. Considering that M-LLM might generate hallucination contents as noises, we use consistency training to reduce noise effect.

3.2. Instructed Multimodal Relation

Since most phrases and their images are not equivalent to each other in MXPR data resource, it is hard to align the two modalities. We bypass

Figure 2: Examples of *equivalent*, *related*, *unrelated* [3] phrase-image pairs with their relation texts generated by mPLUG-Owl.

the alignment issue by directly instructing M-LLMs to output the relations between the phrases and their images. M-LLMs inherit from LLMs the strong in-context learning ability, which enables LLMs to perform tasks given a task instruction or example inputs. In particular, we instruct mPLUG-Owl[3] (Ye et al., 2023), which is an instruction-tuned competitive M-LLM, to output the multimodal relations given a phrase-image pair input. As illustrated in Figure 1, mPLUG-Owl is instructed to answer the question "How is this image related to [phrase]?", where [phrase] is the template slot for filling concrete phrases.

Figure 2 lists three answer examples for *equivalent*, *related*, and *unrelated* phrase-image pairs, respectively. It shows that mPLUG-Owl captures the relation for *equivalent* and *related* with reasonable outputs of relation depiction. Regarding *unrelated* cases, mPLUG-Owl is forced to answer although no relation exists. Since *unrelated* appears very rarely as shown in Table 2, its inaccurate answers do not influence the training. In the end, the overall performance is improved given the instructed multimodal relations as shown in the experiment section 5.1.

3.3. Training

After we get the relation text for a phrase-image pair, we insert it into the example sentence pool for the phrase. The original example sentence pool is provided by WikiXPR (Zheng et al., 2022), which uses sentence level representations of example sentences to extract phrase representation. The incorporation of relation texts into the pool can enrich the phrase contexts to better express the phrase semantics, leading to better phrase representation. We train our model to close the distance between both side phrase representations for each phrase pair.

Let D be the training set of phrase pairs. Each pair is denoted as (p, q), where p is the source phrase, and q is the target phrase. Let i be the image retrieved by querying either p or q, depending on which one is an English phrase since the image retrieval is conducted per English phrase. In the following, we take the source side language as English for example. The training process is the same for the case of English as the target language.

Given p, the example sentence x containing p, and v expressing the relation between p and i, we compose a new example sentence $\bar{x} = [x; eos; v]$, where eos is the delimiter, and ';' denotes the concatenation operator. We feed $\bar{x}$ to MXPR encoder as shown in Figure 1 to get representations of all words, and average the representations of words in p to get phrase level representation $\bar{P}$. If there are multiple example sentences, we repeat the above process, and average all representations of p into the final representation $\bar{P}$. Regarding q, i does not participate in the computation of the representation of q. Only the original example sentences containing q are used to get the phrase representation Q. Both $\bar{P}$ and Q are further updated by a projection head consisting of two linear layers with a ReLU in between and an l_2 normalization

[3] https://github.com/X-PLUG/mPLUG-Owl

[3] In the *unrelated* example, Albanian First Division is a football domestic league competition, while the image is unrelated, resulting in incorrect relation text.

followed.

Multimodal Loss We set a multimodal loss for aligning $\bar{P}$ and Q:

$$\begin{aligned} \mathsf{L}_{\text{multi}}(\bar{P} \rightarrow Q) &= -\log Prob(\bar{P} \rightarrow Q) \\ &= -\log \frac{\exp(\frac{\bar{P} \cdot Q}{t})}{\sum_{Q_j \in B} \exp \frac{\bar{P} \cdot Q_j}{t}} \end{aligned} \quad (1)$$

where B is the training batch containing p and q, $Prob(\bar{P} \rightarrow Q)$ denotes the probability of $\bar{P}$ aligned to Q in B, '$\cdot$' denotes the dot production, and t is the temperature. It is a contrastive loss that drives $\bar{P}$ and Q closer, while separates negative examples further. We compute the overall multimodal loss bidirectionally: $\mathsf{L}_{\text{multi}} = \mathsf{L}_{\text{multi}}(\bar{P} \rightarrow Q) + \mathsf{L}_{\text{multi}}(\bar{P} \leftarrow Q)$.

Consistency Training To alleviate the influence of incorrect relation texts generated by mPLUG-Owl, we build a consistency training scheme to mutually learn between textual and multimodal informations. When there is hallucinatory noise in multimodal relation, the model effectively resorts to textual information. Conversely, when multimodal relation contains valid information, the model utilize it to enrich the phrase representation besides using textual information.

Consistency training is achieved by minimizing the gap between the probability distributions of multimodal prediction and textual prediction. We approximate the distributions in a mini-batch mode. Given p and q in a mini-batch B, the probability of the multimodal prediction is $Prob(\bar{P} \rightarrow Q)$ in equation 1, and the probability of the textual prediction is $Prob(P \rightarrow Q)$. Consistency training is to reduce the Kullback-Leibler (KL) divergence between these two distributions:

$$\mathsf{L}_{\text{con}}(\bar{P} \rightarrow Q) = \mathsf{KL}(Prob(\bar{P} \rightarrow Q) || Prob(P \rightarrow Q)) \quad (2)$$

We also compute the overall consistency loss bidirectionally: $\mathsf{L}_{\text{con}} = \mathsf{L}_{\text{con}}(\bar{P} \rightarrow Q) + \mathsf{L}_{\text{con}}(\bar{P} \leftarrow Q)$.

Overall Training Objectives Besides the multimodal loss and the consistency loss, we also add the original textual loss L_{text} used in WikiXPR. The overall loss is: $\mathcal{L} = \mathsf{L}_{\text{multi}} + \mathsf{L}_{\text{con}} + \mathsf{L}_{\text{text}}$, and is summed over the training set.

3.4. Inference

After training, all phrases obtain their phrase representations. During testing, given a source phrase and its representation in the test set, we retrieve its translation by finding the neighbor closest to its representation vector. The distance function between the source representation and the target representation is the cosine similarity used in the training.

4. Experiments

We use MXPR data resource for training, validation, and testing. Its textual part is the same to WikiXPR (Zheng et al., 2022) for fair comparison. Three tasks are experimented to evaluate the effectiveness of the multimodal information:

- Bilingual phrase retrieval: Given a set of phrases in the source language and a set of phrases in the target language, with optional related images, the task is to find the parallel phrases among them. One model is trained for one language pair in this setting.
- Multilingual phrase retrieval: Given multiple sets of phrases, which are in different languages, with optional related images, the task is to find the parallel phrases among them. One model is trained for all language pairs.
- Zero-shot transfer: One model is trained on one language pair, and is tested on another language pair.

4.1. Experimental Configuration

For generating relationships between images and phrases, we use the open-source mPLUG-Owl model, which is a pre-trained M-LLM, and we utilize the version fine-tuned through instruction tuning. In our experiments, we use sampling with a top-k value of 5 to generate relationships between images and phrases, allowing for the generation of a maximum of 128 words.

For fair comparison to WikiXPR (Zheng et al., 2022), we employed the same XLM-R_{base} (Conneau et al., 2020) to initialize MXPR encoder. During training, we use the same batch size of 256 phrase pairs, and use four example sentences and one optional multimodal relation text for each phrase. MXPR model is trained for 100 epochs with a learning rate of 2×10^{-5}, with 1% warm-up steps and a linear decay throughout the training process. The temperature t is tuned on the validation set. In the case that a phrase does not have a retrieved image, we simply copy the example sentence to replace the relation text to constitute $\bar{x}$.

4.2. Results

Table 3 lists the test set results of the comparison between our multimodal approach MXPR and the textual approach WikiXPR on the three evaluation tasks. Results are averaged over three random seeds in both the xx→en and en→xx directions, where 'xx' denotes one of the eight non-English languages. We also include the strong engine of Google Translate[4] for comparison. Regarding

[4]The results are obtained on January 25, 2024 via https://translate.google.com/.

Table 3: Accuracy@1 of the cross-lingual phrase retrieval on the three tasks. * denotes the statistical significance ($p < 0.01$, using t-test) of the difference between the performances of MXPR and the corresponding WikiXPR.

	En-Fr	En-Ar	En-De	En-Es	En-Ko	En-Ru	En-Zh	En-Ja	Avg
Google Translation	39.51	53.70	32.30	39.78	38.64	45.67	38.02	42.71	41.45
Bilingual phrase retrieval									
WikiXPR	80.18	88.63	81.44	84.53	80.83	91.00	77.62	87.32	83.94
MXPR	**81.39***	**90.22***	**84.08***	**85.95***	**83.73***	**92.11***	**80.36***	**88.18***	**85.75***
Multilingual phrase retrieval									
WikiXPR	85.16	91.90	82.76	90.79	88.22	93.09	**86.47**	90.16	88.56
MXPR	**88.58***	**93.07***	**86.02***	**91.46**	**90.95***	**94.37***	86.14	**91.36***	**90.24***
Zero-shot transfer									
WikiXPR	**77.36**	74.12	73.60	**82.54**	77.91	78.52	77.62	73.04	76.99
MXPR	76.48	**76.72***	**75.46***	82.47	**79.45***	**80.32***	**80.36***	**76.66***	**78.49***

the zero-shot transfer task, we follow the previous work (Zheng et al., 2022) to train the model on the En-Zh training set and subsequently test on other language pairs.

Firstly, the generative translation engine of Google performs significantly worse than WikiXPR and our MXPR. It indicates that the generative translation model is not competent in phrase level translation. Even in high resource language pairs such as English-French, Google Translation only achieves an accuracy of 46.12, which is inferior to 88.58 of our MXPR. In comparison, cross-lingual phrase retrieval approaches perform much better than the generative translation model regarding the phrase level accuracy.

Secondly, we observe that across all the three tasks, MXPR consistently outperforms WikiXPR with an average improvement of 1-2%, achieving state-of-the-art performance on this dataset. This underscores the effectiveness of incorporating multimodal information in cross-lingual phrase retrieval. In the multilingual phrase retrieval task, MXPR achieves the best results in multiple language directions, attaining an average accuracy of 90.24. Compared to the bilingual task, the multilingual task enables MXPR to leverage supervisory signals from other language directions to enhance the model performance. MXPR also shows better generalization ability in transferring the knowledge of the parent MXPR model to other language pairs that have no training data via zero-shot transfer.

4.3. Ablation Study

We conduct ablation studies by removing main components from MXPR in the bilingual phrase retrieval task. In particular, we compare three variants of MXPR that are trained without consistency loss, textual loss, or multimodal loss.

Table 4 shows the result. There is a noticeable performance drop when the consistency loss L_{con} is removed in training. This indicates that consistency loss is important for the training. It can effectively balance the information gain of the textual and image modalities, and reduce the noise effect of the hallucinations generated by M-LLM as manifested in the analysis of section 5.3. Furthermore, if we remove the consistency loss L_{con} and the textual loss L_{text} together, only the multimodal loss L_{multi} is kept for training. It shows that training L_{multi} alone leads to significant performance improvement over WikiXPR when compared to Table 3. Meanwhile, MXPR-L_{con}-L_{text} performs similar to MXPR-L_{con}, demonstrating that L_{text} contribute marginally to the overall improvement. Finally, we remove L_{con} and L_{multi} together, only L_{text} is saved for training, which is equivalent to rerunning WikiXPR. We can see that both L_{con} and L_{multi} contribute most to the overall performance improvement. In summary, this study highlights the efficacy of our proposed multimodal loss and consistency training in improving the cross-lingual phrase retrieval performance.

5. Analyses

5.1. Comparison between Different Categories of the Multimodal Relation

The relation of a phrase-image pair is divided into three categories, i.e., *equivalent*, *related*, and *unrelated*, as listed in Table 2. MXPR exhibits different performances for different categories. Table 5 shows the differences with the comparison between WikiXPR and MXPR. In particular, MXPR is tested with phrase pairs coupled with the related images, while WikiXPR is tested on the same phrase pairs only.

It shows that for the phrase pairs with the relations of *equivalent* and *related*, MXPR performs significantly better than WikiXPR. Since both relations indicate that the phrases and images are highly correlated, their relation texts generated by mPLUG-Owl contain valuable information for the phrase representation, leading to the significant

Table 4: Ablation results of MXPR on the bilingual phrase retrieval task. * denotes the statistical significance ($p < 0.01$, using t-test) of the difference between the performances of MXPR and its ablated versions.

	En-Fr	En-Ar	En-De	En-Es	En-Ko	En-Ru	En-Zh	En-Ja	Avg
MXPR	81.39	**90.22**	**84.08**	**85.95**	**83.73**	**92.11**	**80.36**	**88.18**	**85.75**
MXPR-L_{con}	**81.50**	88.16*	82.40*	85.50	81.27*	91.93	78.43*	87.80	84.62*
MXPR-L_{con}-L_{text}	81.39	88.77*	83.07*	85.73	81.13*	91.88	78.74*	87.25*	84.74*
MXPR-L_{con}-L_{multi}	80.01*	88.71*	81.86*	84.83*	80.81*	91.12*	78.00*	87.43*	84.09*

accuracy improvement. However, for phrase pairs with the relation of *unrelated*, there is a decrease in accuracy for En-De and En-Fr. This suggests that unrelated images may introduce noise for the phrase representation. For the other language pairs, MXPR performs similar to WikiXPR in the cases of unrelated images. Due to the tiny portion of the relation of *unrelated*, its effect is marginal to influence the overall performances. The improved performances of the majority relations of *equivalent* and *related* proves the quality of the created MXPR data resource.

5.2. Effect of the Multimodal Relations

Since the multimodal relation text is inserted into the example sentence pool in MXPR, the performance improvement of MXPR over WikiXPR may be thought as the outcome of the enlarged example sentence pool. To study whether the reason of the improvement lies in the enlarged pool, we keep the size of the pool unchanged by randomly deleting one example sentence in the original pool and inserting the relation text generated by mPLUG-Owl into the pool. The results are listed in table 6. $MXPR_{del}$ denotes this study.

Table 6 shows that $MXPR_{del}$ performs similar to MXPR, still significantly surpassing WikiXPR. This indicates that the improvement of MXPR is not merely due to adding sentences but rather comes from leveraging multimodal information.

Furthermore, we also test the effect of the multimodal relations by using random images as the 'related' images for phrases. $MXPR_{rand\text{-}image}$ denotes this process. It shows that the performance is worse than WikiXPR due to the noise of the random images, while MXPR is significantly better than WikiXPR due to the utilization of the original images. This result states again the importance of the multimodal information in improving the performance.

5.3. Qualitative Analysis of Consistency Training

Consistency training tries to balance the contributions from image and text, making the two modalities complement to each other. Figure 3 presents two qualitative examples from the test set of En-Zh for illustrating the effect of consistency training. Both examples list the result of WikiXPR that only uses the textual loss, the result of $MXPR_{multi}$ that only uses the multimodal loss, and the result of MXPR that uses consistency training to balance the two modalities.

In the first example, given the input phrase 'Cantonese Opera', WikiXPR result is '中央芭蕾舞团' ('National Ballet of China'), which is incorrect. This kind of error is likely due to the similarity in the contexts of the two phrases in the pure textual modality. With the addition of the multimodal relation text, $MXPR_{multi}$ is able to leverage cues such as 'traditional costumes' to correctly identify the answer '粵劇', indicating that the image provides information complementary to the text. Through consistency training, MXPR resorts to the multimodal relation text to get correct prediction.

However, due to the presence of hallucinations in M-LLMs, the multimodal relation texts are not always beneficial for the cross-lingual phrase retrieval. In the second example, given the input phrase 'Order of the Sacred Treasure', which is a kind of Japanese medal, and its related image, the multimodal relation text generated by M-LLMs is filled with unrelated information to the phrase, and misleads $MXPR_{multi}$ to predict the wrong result '世界遺产委員會' ('World Heritage Committee'). It is worth noting that after consistency training, MXPR resorts to the textual modality, being consistent with WikiXPR and rectifying the error caused by the multimodal relation text.

5.4. Visualization

Figure 4 depicts examples of the phrase embeddings trained by MXPR on En-Zh test set. The dimensions of the embeddings are reduced to two by using T-SNE. In particular, we sample English phrases from the test set at first, then search for their nearest Chinese neighbors in the phrase embedding space. It shows that phrases in a translation pair do appear as the closest neighbors in the 2-dimension visualization.

6. Related Works

Cross-lingual phrase retrieval We introduce cross-lingual retrieval and phrase retrieval at first, then we introduce cross-lingual phrase retrieval.

Cantonese Opera

This image depicts a Cantonese opera performance, as two actors are dressed in elaborate traditional costumes, performing on stage. The actors are wearing colorful, ornate outfits and appear to be engaged in an emotional scene during the production…

WikiXPR :中央芭蕾舞团 (National Ballet of China)

MXPR$_{multi}$:粤劇

MXPR :粤劇

Order of the Sacred Treasure

This image features a gold and red medal… and is likely part of the Order of the Sacred Treasure (OOTS), a fictional organization from the webcomic and video game series by web artist Randall Munroe. OOTS is an organization that seeks to protect and preserve the world… The snowflake-like design of the medal is representative of the organization's commitment to preserving nature and the world…

WikiXPR :瑞寶章

MXPR$_{multi}$:世界遺产委員會 (World Heritage Committee)

MXPR :瑞寶章

Figure 3: Examples from from En-Zh test set. Input phrases and their related images are shown in the left part. The multimodal relation text generated by mPLUG-Owl is in the middle. Retrieval results are listed in the right.

Table 5: Test set accuracy@1 of WikiXPR and MXPR on the bilingual phrase retrieval task across the three categories of the multimodal relation.

	En-Fr		En-Ar		En-De		En-Es		En-Ko		En-Ru	
	WikiXPR	MXPR	WikiXPR	MXPR	WikiXPR	MXPR	WikiXPR	MXPR	WikiXPR	MXPR	WikiXPR	MXPR
equivalent	78.82	80.00	89.77	90.19	72.86	75.58	81.54	81.54	73.64	78.29	92.22	93.46
related	80.37	81.97	87.20	90.23	82.42	86.24	85.97	87.68	83.18	85.39	91.04	91.51
unrelated	71.42	64.28	91.99	91.99	88.23	85.29	62.50	62.50	73.07	73.07	97.82	97.82

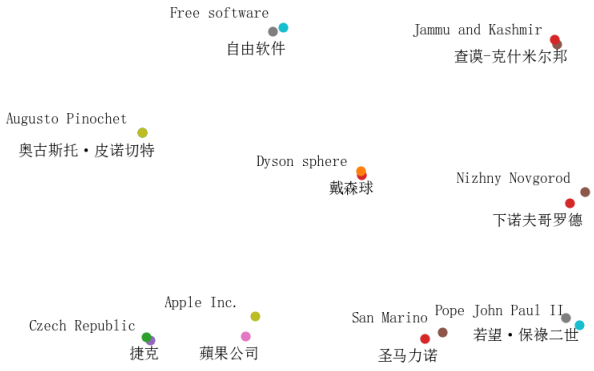

Figure 4: Visualization of the bilingual phrase embeddings trained by MXPR.

We state the difference to our work at last.

Cross-lingual retrieval can be categorized into word-level and sentence-level approaches. Word-level approaches typically involve training word embeddings separately for each language and aligning these embeddings across languages based on a learned mapping function (Mikolov et al., 2013; Artetxe et al., 2016; Joulin et al., 2018; Doval et al., 2018). Sentence-level approaches learn language-agnostic sentence representations, which enables sentences to be retrieved across languages (Conneau and Lample, 2019; Artetxe and Schwenk, 2019; Conneau et al., 2020; Goswami et al., 2021).

Phrase retrieval involves learning phrase representations to retrieve relevant phrases. In monolingual settings, phrase retrieval has shown significant success in tasks such as open-domain question answering (Seo et al., 2019) and entity linking (Gillick et al., 2019; Lee et al., 2021). Zheng et al. (2022) extend phrase retrieval to cross-lingual phrase retrieval, wherein phrase representations are learned using monolingual data and are aligned to retrieve parallel phrases across different languages. In contrast, our work is the first to utilize multimodal information for cross-lingual phrase retrieval.

Multimodal language processing Inspired by pre-trained large language models (Brown et al., 2020), visual-language pre-training models have also made significant strides through pre-training on large-scale image-text pairs (Li et al., 2020a) for the cross-modal tasks such as VQA (Zhou et al., 2020) and image-text retrieval (Radford et al., 2021). The emergence of multimodal large language models has further enhanced contextual learning capabilities for cross-modal tasks (Ye et al., 2023). Li et al. (2020c) suggest that the addition of object detection labels can improve the cross-modal capabilities in multimodal pre-training.

Table 6: Accuracy@1 of the test sets for the multimodal relation comparison study. * denotes the statistical significance ($p < 0.01$, using t-test) of the difference between the performances of MXPR and its variants or WikiXPR.

	En-Fr	En-Ar	En-De	En-Es	En-Ko	En-Ru	En-Zh	En-Ja	Avg
MXPR	**81.39**	**90.22**	**84.08**	**85.95**	**83.73**	**92.11**	**80.36**	**88.18**	**85.75**
$MXPR_{del}$	80.93	90.12	83.92	85.84	83.38	**92.11**	80.27	87.91	85.56
$MXPR_{rand\text{-}image}$	79.56*	88.20*	80.04*	83.70*	79.73*	91.10*	77.80*	87.18*	83.41*
WikiXPR	80.18*	88.63*	81.44*	84.53*	80.83*	91.00*	77.62*	87.32*	83.94*

Regarding multilingual multimodal pretraining and its applications, Ni et al. (2021) propose M3P, which is a multitask multilingual multimodal pretrained model, and achieves excellent results in multilingual cross-modal tasks such as multilingual image-text retrieval. UC2 unifies cross-lingual cross-modal pre-training through techniques such as data augmentation via machine translation, visual context as pivot, and multitasking (Zhou et al., 2021). MURAL is pretrained on large-scale multilingual image-text pairs and bilingual translation pairs from the web using contrastive objectives for both image-text and text-text tasks (Jain et al., 2021). Different to the above multimodal sentence-level or long text tasks, our task is to learn cross-lingual phrase-level representations enriched by multimodal information.

7. Conclusion

Current cross-lingual phrase retrieval approaches only deal with textual modality. In this paper, we introduce multimodal information into the task by creating the first multimodal cross-lingual phrase retrieval data resource and building a framework based on the data resource. The framework bridges the gap between different modalities by using multimodal relation generation and consistency training. The resultant phrase representations are enriched by the multimodal information, achieving significant improvement over the pure textual approaches across various language pairs in extensive experiments.

Acknowledgements

We would like to thank the anonymous reviewers for the helpful comments. This work was supported by National Natural Science Foundation of China (Grant No. 62261160648, 62276179) and Alibaba Innovative Research Program.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2289–2294.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North*, pages 4159–4170. 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. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 294–304.

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.

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John Philip McCrae. 2021. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomáš Mikolov, Hervé Jégou, and Édouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2021. 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.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11336–11344.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Juntao Li, Chang Liu, Jian Wang, Lidong Bing, Hongsong Li, Xiaozhong Liu, Dongyan Zhao, and Rui Yan. 2020b. Cross-lingual low-resource set-to-description retrieval for global e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8212–8219.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. 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 852–860.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3976–3985.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Andreas Rücklé, Krishnkant Swarnkar, and Iryna Gurevych. 2019. Improved cross-lingual question retrieval for community question answering. In *The world wide web conference*, pages 3179–3186.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4430–4441, Florence, Italy. Association for Computational Linguistics.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4346–4350.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Heqi Zheng, Xiao Zhang, Zewen Chi, He-Yan 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.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4153–4163.