IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 18, 2020, accepted November 27, 2020, date of publication December 1, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041605

Cross-Lingual Passage Re-Ranking With Alignment Augmented Multilingual BERT

DONGMEI CHEN, SHENG ZHANG, XIN ZHANG, AND KAIJING YANG
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China
Corresponding author : Sheng Zhang (zhangsheng@nudt.edu.cn)

ABSTRACT The task of Cross-lingual Passage Re-ranking (XPR) aims to rank a list of candidate passages in multiple languages given a query, which is generally challenged by two main issues: (1) the query and passages to be ranked are often in different languages, which requires strong cross-lingual alignment, and (2) the lack of annotated data for model training and evaluation. In this article, we propose a two-stage approach to address these issues. At the first stage, we introduce the task of Cross-lingual Paraphrase Identification (XPI) as an extra pre-training to augment the alignment by leveraging a large unsupervised parallel corpus. This task aims to identify whether two sentences, which may be from different languages, have the same meaning. At the second stage, we introduce and compare three effective strategies for cross-lingual training. To verify the effectiveness of our method, we construct an XPR dataset by assembling and modifying two monolingual datasets. Experimental results show that our augmented pre-training contributes significantly to the XPR task. Besides, we directly transfer the trained model to test on out-domain data which are constructed by modifying three multi-lingual Question Answering (QA) datasets. The results demonstrate the cross-domain robustness of the proposed approach.

INDEX TERMS Passage re-ranking, cross-lingual learning, pre-training tasks.

I. INTRODUCTION

Passage re-ranking is an essential task in many Natural Language Processing (NLP) applications such as passage retrieval for open-domain question answering. It requires a system to rank a list of candidate passages based on the provided query. A number of approaches based on neural networks have been proposed to perform this task, e.g., KNRM [1], DUET [2], Co-PACRR [3] and BERT [4], among which the BERT-based techniques are shown to achieve superior performance.

In the existing passage re-ranking literature, it is commonly assumed that the query and the passages to be ranked are both in the same language, e.g., English or Chinese. It is then learned to rank the passages by scoring the semantic similarity between each of them and the query. However, the monolingual assumption does not always uphold. For instance, suppose a scenario in which a Chinese immigrant with limited command of English buys a car and then asks the "salesman", a service robot, a question in Chinese "我可以在 没有保险的情况下开新车回家吗?" (can I drive the new car home without insurance?), while the relevant passage "Generally speaking, probably not. Most States require you..." is in English. Here, a passage re-ranking module is required for the robot which can perform passage re-ranking in a cross-lingual scenario. To address this issue, in this article we explore Cross-lingual Passage Re-ranking (XPR), which refers to *ranking a list of candidate passages in multiple languages, of which only a portion are in the same language as the query*.

The XPR is a challenging task due to at least two reasons. The first is that the query and passages are often in different languages. Especially, languages that belong to different language families, such as English and Chinese, have different orderings. Thus, it is difficult to evaluate the relevance between a query in one language, say Chinese, and the passages to be ranked in another language, say English. The second issue for the XPR is lack of annotated data which is indispensable for model training and performance evaluation. However, labeling a sufficient amount of data is also resource-demanding and time-consuming.

To accomplish this challenging task, inspired by the impressive performance of BERT-based models for monolingual passage re-ranking, we choose to utilize its multilingual

The associate editor coordinating the review of this manuscript and approving it for publication was Wenge Rong.

extension, i.e., multilingual BERT (mBERT) [4]. mBERT has been trained on a large set of Wikipedia data in 104 languages. Nevertheless, this training data contains no explicitly parallel sentences, i.e., translated pairs, therefore the resulting representations lack alignment between languages. This is however very important for addressing the first challenge of the XPR task mentioned above. To augment the cross-lingual alignment, here we introduce Cross-lingual Paraphrase Identification (XPI) as an extra pre-training task. The XPI is to identify whether two sentences from different languages have the same meaning. To address the second issue, i.e., to overcome the shortage of annotated data, we modify an existing monolingual dataset with its translation to construct the training and testing datasets. To construct the training dataset, we present three strategies, i.e., merging, cascading, and mixing. For the testing dataset, we randomly replace half of the queries with their translations in a coin-tossing manner and replace half of the passages associated with each query with their translations. Besides, in order to verify the robustness of our method, we directly transfer the trained model to test on three real-life-like datasets. Experimental results indicate that our proposed approach substantially boosts the model performance on the XPR task.

The main contributions of this article are as the following.

- To the best of our knowledge, this is the first work exploring the generic task of cross-lingual passage re-ranking, which aims to rank a list of candidate passages with respect to a given query. It is ‘‘generic’’ due to (1) different queries are not restricted to be in one particular language, and (2) at least a portion, if not all, of the passages to be ranked for each query are in languages different from that of the query.
- To augment the alignment between the representations of different languages, we introduce an extra pre-training task, viz. XPI, that aims to identify whether two sentences, which may be from different languages, have the same meaning.
- To alleviate the scarcity of annotated data for XPR, we create a new dataset by combining a monolingual one with its already available translation. Besides, we present three effective strategies for model training. Extensive experiments have been conducted on in-domain as well as out-domain sets. The results prove the effectiveness and robustness of our proposed method.

The remainder of this article is structured as follows. Section II provides a review of the related work to provide the context for our contributions. Section III describes and formalizes the task of XPR. Then in Section IV, we introduce our method followed by the details of the experiments in Section V. The experiment results and analysis are presented in Section VI. We finally conclude the article and highlight interesting research directions for future works in Section VII.

II. RELATED WORK

In the following, we briefly review three broad classes of the related works, including, neural ranking models, cross-lingual learning, and pre-training tasks.

A. NEURAL RANKING MODELS

Neural ranking models are categorized into *representation-focuse* and *interaction-focused* families [5]. The former family of models, e.g., DSSM [6], CDSMM [7] and ARC-I [8], compute a set of semantic representations of the query and the passage and then use a set of simple functions to evaluate the final relevance score. These approaches explicitly model the process of matching the query and the candidate passages. Nevertheless, these methods do not usually consider the interaction between them. This is deemed to be helpful to avoid the impact of the parts in the document irrelevant to the query. In contrast, the latter family of models, e.g., ARC-II [8], KNRM [1] and CONV-KNRM [9], directly define the interaction functions and use a set of complex evaluation functions to abstract the interaction and compute the relevance score. BERT-based models, as pre-trained models, enjoy the merits of both architectures and are shown to have a surprisingly high performance on various NLP tasks. Researchers have begun to apply them to the tasks of monolingual passage re-ranking. For instance, Nogueira *et al.* [10] describe a BERT-based model and prove the effectiveness of the model for the task of passage re-ranking. Yang *et al.* [11] further explore the applications of BERT to ad hoc document retrieval. It is noteworthy that all of the above works focus on monolingual (mostly English) passage ranking.

Also, a growing number of researches are dealing with cross-lingual passage ranking. Hull *et al.* [12] propose a task called Cross-lingual Information Retrieval (CLIR). Given a query in one language, the task is to retrieve relevant documents in another language. More recently, Martino *et al.* [13] study how to find relevant questions in community forums when the language of the given question is different from that of the candidate questions. The dataset of BUCC [14] and Tatoeba [15] consider the task of, given an English sentence, retrieving the parallel sentence from a monolingual pool of candidates in another language. While these studies are inherently cross-lingual, they limit the input queries (e.g., sentences or questions) to be all in the same language, and the candidates to be in another language different from that of the query. Two latest works, i.e., LAReQA [16] and XOR QA [17], focus on cross-lingual question answering and are more closely related to ours. Specifically, both of them eliminate the limitation that requires the candidates to be in the same language, which is similar to our task settings. However, XOR QA restricts the questions to be monolingual, and hence can be regarded as a special case of the XPR task we are addressing. In addition, LAReQA creates a set of parallel pairs for each question-answer in a particular language, say English, via manual translation. Then, given a question in some language, the task is to retrieve all its answers including the original one as well as the parallel ones with the goal of assessing a model's ability to achieve cross-lingual alignment. In comparison, the XPR task we focus on aims to support real-world cross-lingual applications such as intelligent customer service described in Section I. Hence, we believe the settings are more practically oriented.

B. CROSS-LINGUAL LEARNING

The diversity of languages is a major challenge in NLP. Annotating large quantities of data for each language is almost impossible unrealistic. An alternative is to leverage cross-lingual learning, which is discussed in this section. It is essentially a transfer learning problem across different languages. Mikolov *et al.* [18] first leverage a bilingual dictionary to align the word representations in different languages. Further research in this area, resulted in a further reduction of dependency on bilingual dictionaries. Gouws *et al.* [19] proposed a method in which bilingual word representations are learned jointly from parallel corpora without word alignment. Furthermore, Conneau *et al.* [20] propose an unsupervised learning method that obtains cross-lingual word representations without dictionaries or parallel data. Very recently, there has been a surge of interest in leveraging cross-lingual pre-trained models, such as mBERT [4] and XLM [21]. These models are trained using multilingual datasets and achieve state-of-the-art performance in many cross-lingual NLP tasks [22] except for cross-lingual passage re-ranking. The existing works essentially adopt transfer learning. In other words, a model is first trained based on languages for which datasets are available. The trained model is then directly transferred to the targeted language for which usually only a limited amount of training and test data is available. Different from these works, our proposed method is concerned with a cross-lingual situation, where the query from one language is applied to ranking passages from multiple languages.

C. PRE-TRAINING TASKS

Inspired by the superior performance of BERT endowed by pre-training, many pieces of research efforts were made towards adding or modifying the pre-training tasks. For instance, Yang *et al.* [23] propose a generalized autoregressive pre-training method to learn bidirectional contexts in a pre-trained language model, XLNet. Furthermore, Sun *et al.* [24] construct seven pre-training tasks covering different aspects of languages to train a transformer-based language model, ERNIE 2.0. Other researchers consider adding or modifying pre-training tasks for multilingual/cross-lingual language models to improve their performance. For instance, mBERT [4] is trained on two unsupervised tasks without using any parallel data including Masked Language Model (MLM), and Next Sentence Prediction (NSP). XLM [21] is trained on three tasks: two unsupervised tasks that require monolingual data, and a supervised task that requires parallel data. These three tasks of XLM are Masked Language Model (MLM), Causal Language Model (CLM), and Translation Language Model (TLM), respectively.

Nevertheless, since in training of these cross-lingual pre-trained models explicitly parallel sentences have not (or only partially) been used, the resulting representations lack alignment between languages, especially in the sentence-level. Sentence level alignment is essential for our task because the query and passages to be ranked are often in different languages. Therefore, we propose a sentence-level pre-training task to align the language representations.

III. TASK DEFINITION

The XPR task we consider in this article is described as the following. Given a query such as "Can I drive a new car home without insurance?" or "我可以在没有保险的情况下开新车回家吗?" (Can I drive a new car home without insurance?), the system is required to re-rank a list of candidate passages in multiple languages, where each passage is typically in a single language.

For simplicity of notation and without loss of generality, we focus on a bilingual case, where the two languages are denoted as l_1 and l_2, respectively. We are given a query, q, that is either in l_1 or l_2 along with a list of candidate passages, $P = [p_1^{l_1}, p_2^{l_1}, \ldots p_m^{l_1}, p_{m+1}^{l_2}, p_{m+2}^{l_2}, \ldots p_{m+n}^{l_2}]$, which consists of m passages in l_1 and n passages in l_2. We assume that there is a list of passages, P_q^*, which are truly related to q, i.e., P_q^* is a ground-truth list. The objective of XPR is to construct an evaluation function, $S(q, p_i)$, to score each query-passage pair, (q, p_i) as $S(q, p_i)$. The function needs to ensure that a higher $S(q, p_i)$ is associated with a higher likelihood of p_i belonging to P_q^*.

In this article, we consider English and Chinese as the examples for describing our methodology and presenting our experiments. It is worth noting that our method is not limited to these two languages and easy to be used for other languages.

IV. METHODOLOGY

A. OVERVIEW

The proposed framework for XPR is illustrated in Fig. 1, where the upper row indicates the training process, and the lower row shows the models involved in each stage of the training process. It is based on the "pre-training+fine-tuning" architecture. The training process consists of two stages. The first stage is *extended pre-training*, which is proposed in this article as an extra pre-training task to augment the cross-lingual alignment. The input to this stage is the model shown in Fig. 1(b), which contains Input, mBERT, XPI Head, and Output. The weights of mBERT are initialized using those pre-trained and released by Devlin *et al.* [4]. The resulting model of stage one, see, Fig. 1(c), is further used for constructing the input model of the next stage. We establish the input to stage two, i.e., *target task-oriented fine-tuning*, by replacing the task head of model in Fig. 1(c) with the XPR head which results in the model shown in Fig. 1(d). After stage two, the model in Fig. 1(e) together with trained weights

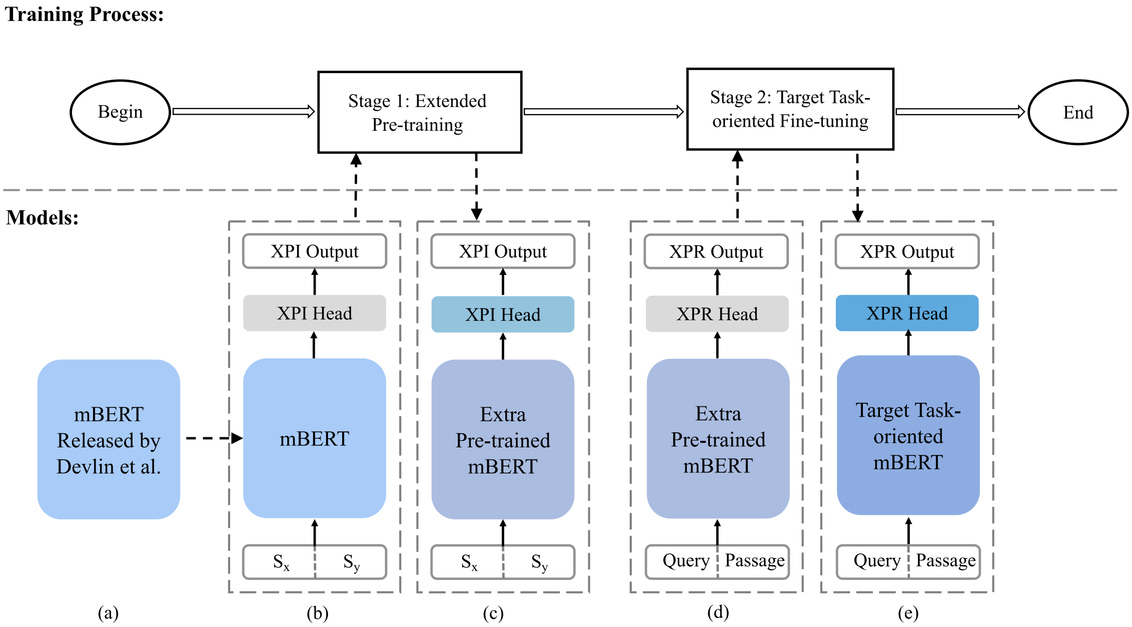

FIGURE 1. The framework of cross-lingual passage re-ranking. Different colors indicate different model parameters, e.g., grey color denotes random initialization. [Best viewed in color.]

are then used for target task at the test time. In the following, we elaborate on these two training stages.

B. EXTENDED PRE-TRAINING

As mentioned above, mBERT has the limitation that it is pre-trained based on monolingual corpora in 104 languages without explicitly using any parallel data. Therefore, its performance for cross-lingual tasks is not always high. For instance, mBERT is less accurate in performing cross-lingual tasks involving Chinese and English [25]. A possible explanation for this is the fundamental dissimilarity of the linguistic typology of these two languages. In other words, English and Chinese have different orders of subject, verb, and object.

To address this issue, in this article, we propose an extra pre-training task that incorporates the explicit cross-lingual signals into the training process to augment the alignment across different languages. We further note that different ordering of the sentence elements is reflected at the sentence-level rather than the word-level. Therefore, we propose to use a sentence-level task, i.e., XPI, as an extra pre-training task, which is detailed in the following.

The traditional paraphrase identification (PI) task is to determine whether two monolingual sentences have the same meaning. As our goal is to augment cross-lingual alignment, we modify the PI task by adding cross-lingual sentence pairs into its input. Hence, the task of XPI takes two sentences which may be from different languages as input and identifies whether they have the same meaning.

1) THE PROPOSED MODEL

As mentioned above, we propose an mBERT-based model to perform the XPI task as illustrated in Fig. 2. The model consists of four blocks, including Input, mBERT, XPI Head and Output. As it is seen, the input sequence to mBERT is a concatenation of two sentences, x, and y, separated by special delimiter markers, e.g. [CLS] x [SEP] y [SEP]. For each token of the sequence, its input embeddings are constructed by summing the token embeddings, the segmentation embeddings, and the position embeddings [4]. We then transform the input embeddings into the contextual representations using Transformer Encoder [26] which generates the hidden representation, h_{CLS}, in the final layer. Finally, h_{CLS} is fed into a fully connected layer followed by a sigmoid function. The similarity of two sentences is then formally evaluated as:

$$z_p = sigmoid(W_p h_{CLS} + b_p), \tag{1}$$

where W_p and b_p are trainable task-specific parameters.

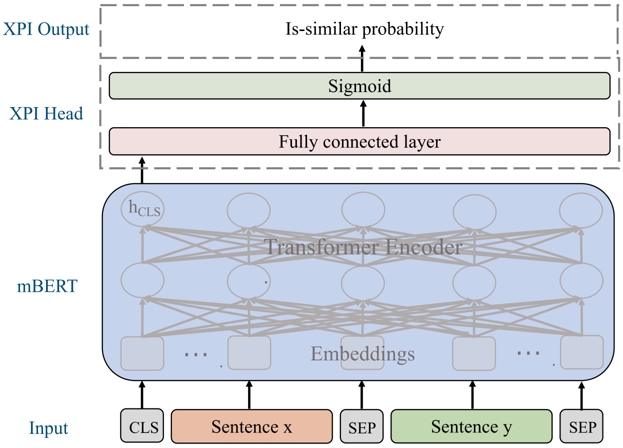

FIGURE 2. mBERT for cross-lingual paraphrase identification.

TABLE 1. The four possible forms of a sentence pair to be fed into the cross-lingual paraphrase identification.

Type	Cross-lingual Paraphrase Identification Instance
English-English	**x:** Deliberate or accidental fires have had disastrous effects on the ecology and economy of the areas concerned. **y:** The impact of disasters on sustainable development has been shown to be very severe.
Chinese-Chinese	**x:** 蓄意纵火或意外火灾已对有关地区的生态和经济产生了灾害性的影响。 **y:** 灾害对持续发展的影响已证明是极其严重的。
English-Chinese	**x:** Deliberate or accidental fires have had disastrous effects on the ecology and economy of the areas concerned. **y:** 灾害对持续发展的影响已证明是极其严重的。
Chinese-English	**x:** 蓄意纵火或意外火灾已对有关地区的生态和经济产生了灾害性的影响。 **y:** The impact of disasters on sustainable development has been shown to be very severe.

2) MODEL TRAINING

To augment the sentence-level cross-lingual alignment, using training, here, we modify the score of similarity defined in Eq.(1) so it is higher for the sentence pairs in different languages with the same meanings. Such a similarity measure can be then utilized as a reasonable criterion for a binary classification. We assume that each input pair is classified into two distinct classes, e.g., labels positive (1) if the sentences within have the same meaning, and negative (0), otherwise. Therefore, the loss function to be minimized in the training is defined as the cross-entropy:

$$L_{XPI} = -\sum_{j=1}^{N}(\hat{z}_p^j log(z_p^j) + (1-\hat{z}_p^j)log(1-z_p^j)), \tag{2}$$

where $\hat{z}_p^j \in \{0, 1\}$ is the ground-truth label of example j, and N is the number of sentence pairs. Instead of training a transformer architecture from the scratch, here we adopt mBERT's pre-trained weights as the initialization for basic language understanding, and then we extend the pre-training with the XPI task.

Motivated by the strategy of data augmentation employed in [27], here we modify some existing parallel data to construct a mix-language dataset for training. In this article, a sentence pair that we refer to as 'mix-language' consists of two sentences in different languages. A sentence pair in the same language is also called 'mono-language'. The mix-language training data is obtained through the following steps: 1) English training data is translated into Chinese; 2) mix-language training data is constructed. To reduce the required manpower for translation, in our implementation we choose a publicly available dataset with its available translation.

We assume that $S^{en} = \{(x_j^{en}, y_j^{en})\}_{j=1}^N$ is an English dataset and $S^{zh} = \{(x_j^{zh}, y_j^{zh})\}_{j=1}^N$ is its Chinese translation, where N denotes the number of sentence pairs in each dataset. We use them to construct a mix-language dataset, S^{mix}. First, we randomly select half of English sentence pairs in S^{en} and then randomly replace half of the selected ones with their Chinese translations. As a result, there are $N/4$ English and $N/4$ Chinese sentence pairs. Then in the second step, for each unselected sentence pair of S^{en} in the first step, we randomly replace x_j^{en} or y_j^{en} in a coin-tossing manner: for a head, we replace x_j^{en} with its Chinese translation, x_j^{zh}, and for a tail, we replace y_j^{en} with y_j^{zh}. A mix-language dataset, S^{mix}, is then obtained. The dataset contains samples from S^{en}, S^{zh} and samples in the form of (x_j^{en}, y_j^{zh}) or (x_j^{zh}, y_j^{en}). Table 1 provides an example for each of the above-mentioned four possible forms.

In summary, there are $N/2$ mono-language and $N/2$ mix-language sentence pairs in S^{mix}. It is noteworthy that for each mix-language sentence pair, either x or y is Chinese. In other words, we only ensure that the two sentences come from different languages. This is to avoid cases where the model incorporates language-specific information into representations through a fixed collocation of languages in each sentence.

C. TARGET TASK-ORIENTED FINE-TUNING

In the following, we first introduce our target task, XPR. Then we describe the three training strategies that are adopted in this article to address the lack of annotated training data.

1) CROSS-LINGUAL PASSAGE RE-RANKING (XPR)

The XPR task aims to estimates how relevant a candidate passage is to a query. Our model for the task is built on top of the extra pre-trained mBERT. Similar to the setup of XPI task, here, we concatenate query, q, and passage, p, as a sequence, [CLS] q [SEP] p [SEP] and feed it as an input to the extra pre-trained mBERT. Then the output representation, $\tilde{h}_{CLS}$, of the Transformer Encoder is fed into a fully connected layer followed by a sigmoid function. Finally, we obtain the relevancy score for the input query and the passage as:

$$z_r = sigmoid(W_r\tilde{h}_{CLS} + b_r), \tag{3}$$

where W_r and b_r are trainable parameters. In the supervised setting, we assume that the ground-truth label for each query-passage pair is binary, i.e., "1" represents relevancy and "0" represents irrelevancy.

To train our re-ranking model, we continue with training the extra pre-trained mBERT, and then fine-tune it using negative log probability. We train all the parameters as well as W_r and b_r by minimizing the sum of the negative log probability of the correct labels. The loss function is defined as:

$$L_{XPR} = -\sum_{i=1}^{M}(\hat{z}_r^i log(z_r^i) + (1-\hat{z}_r^i)log(1-z_r^i)), \tag{4}$$

where $\hat{z}_r^i$ denotes the ground-truth label of example i, and M denotes the number of query-passage pairs. At the test time, for each query, we independently evaluate the score of each

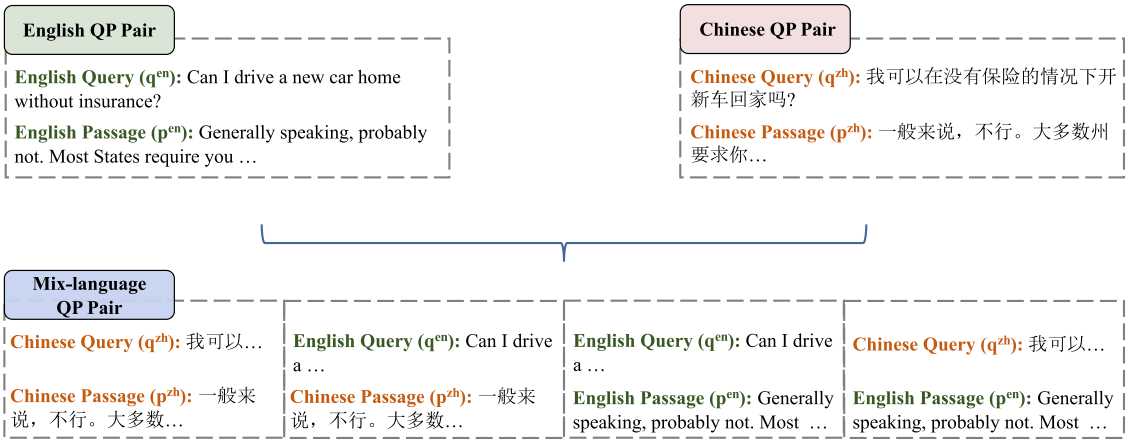
English QP Pair
English Query (q^en): Can I drive a new car home without insurance?
English Passage (p^en): Generally speaking, probably not. Most States require you …
Chinese QP Pair
Chinese Query (q^zh): 我可以在没有保险的情况下开新车回家吗?
Chinese Passage (p^zh): 一般来说，不行。大多数州要求你…
Mix-language QP Pair
Chinese Query (q^zh): 我可以…
Chinese Passage (p^zh): 一般来说，不行。大多数…
English Query (q^en): Can I drive a …
Chinese Passage (p^zh): 一般来说，不行。大多数…
English Query (q^en): Can I drive a …
English Passage (p^en): Generally speaking, probably not. Most …
Chinese Query (q^zh): 我可以…
English Passage (p^en): Generally speaking, probably not. Most …

FIGURE 3. An example of the mixed training strategy. The second row provides possible query-passage pairs. Note that only two of them will be created from the two pairs in the first row, i.e., (q_i^{zh}, p_i^{zh}), (q_i^{en}, p_i^{zh}) or (q_i^{en}, p_i^{en}), (q_i^{zh}, p_i^{en}).

query-passage pair. Finally, we re-rank the list of passages by their resulting scores.

2) TRAINING STRATEGIES

Due to the lack of annotated data for cross-lingual passage re-ranking, we present three training strategies in which we effectively utilize some existing English training data and its Chinese translation. For fair comparisons, the number of training data in different training strategies is equal. We assume $D^{en} = \{(q_i^{en}, p_i^{en})\}_{i=1}^{M}$ to be English training data, and $D^{zh} = \{(q_i^{zh}, p_i^{zh})\}_{i=1}^{M}$ to be its Chinese translation, where M denotes the number of query-passage pairs in each language. In the following, we describe these training strategies in detail.

- Merged Training:

In this training strategy, we simply merge English training data, D^{en}, with its Chinese translation, D^{zh}, such that there are $2M$ samples for training. After obtaining the reconstructed samples, our model is trained from the extra pre-trained mBERT.

- Cascade Training:

This strategy consists of two steps. Firstly, we only train the extra pre-trained mBERT with English training data, D^{en}. Secondly, we continue training with its Chinese translation, D^{zh}, starting from the model which is generated from the first step.

- Mixed Training:

In this training strategy, we modify English training data, D^{en}, with its Chinese translation, D^{zh}, to construct mix-language data D^{mix} for training. Firstly, we randomly select half of English query-passage pairs. For each selected English sample, we then replace p_i^{en} with its translation, p_i^{zh}. For the translation of each selected sample, we keep it in the original language. Therefore, there are $M/2$ mix-language samples in the form of (q_i^{en}, p_i^{zh}), and $M/2$ Chinese samples. Secondly, for each remaining English sample of D^{en} in the first step, we keep it in its original language. For the translation of each remaining sample, we then replace p_i^{zh} with p_i^{en}. After the above two steps, there are M mono-language, and M mix-language samples in D^{mix}. An example of D^{mix} is presented in Fig. 3. Finally, our model is trained from the extra pre-trained mBERT with D^{mix}.

V. EXPERIMENTS SETUP

In this section, we first describe the datasets in Section V-A, then introduce the evaluation metrics in Section V-B, and the baseline in Section V-C. The implementation details are also given in Section V-D.

A. DATASETS

We use three groups of datasets for experimentation. The first contains only the United Nations Parallel Corpus[1] for the training and testing of our proposed XPI model. The second includes solely one dataset to perform training, or rather fine-tuning, and in-domain testing of the XPR model. The last one is comprised of three cross-lingual datasets for the out-domain testing of the trained XPR model to verify its cross-domain robustness.

1) CROSS-LINGUAL PARAPHRASE IDENTIFICATION DATASET

To further align language representations, we propose an extra pre-training task in which we utilize a multilingual parallel corpus, United Nations Parallel Corpus [28]. The reasons why we choose this corpus are as follows. Firstly, the corpus covers a wide variety of domains, e.g., education, economy, etc., and thus can make the model trained on it more universally usable. Secondly, the corpus contains a large number of high-quality parallel sentences provided by human experts, which are beneficial for the training of

[1] https://conferences.unite.un.org/UNCorpus

cross-lingual models. To be more specific, the corpus consists of official records and other parliamentary documents of the United Nations that are available in the public domain. Most of these documents are available in six official languages of the United Nations, including English, Chinese, Spanish, French, Russian, and Arabic. The corpus contains a large number of parallel sentences that were produced and manually translated between 1990 and 2014. The corpus also provides 15,886,041 English Chinese sentence pairs. In this article, we only use 100,000 sentence pairs for the extended pre-training of mBERT.

2) CROSS-LINGUAL PASSAGE RE-RANKING DATASET

Currently, there is no publicly available dataset for XPR. Here, we construct a new dataset by modifying an English dataset, InsuranceQA_v2 [29], together with its translation, i.e., Insuranceqa-corpus-zh.

a: InsuranceQA_v2

It is a well-known passage re-ranking benchmark. The dataset is composed of real word queries from users and passages from experts with domain knowledge of insurance. It contains 20,889 queries in total and has been divided into three parts: training (16,889), development (2,000), and test set (2,000). For each query, 500 candidate passages are retrieved using the SOLR search engine. Thus, there is no guarantee that every associated candidate pool contains a relevant passage. In this article, we discard all queries without any relevant passage in the associated pool. As a result, there are 10391 queries in the training set, 1,592 queries in the development set, and 1,625 queries in the test set.

b: INSURANCEQA-CORPUS-ZH

It is the Chinese translation of InsuranceQA_v2, which is publicly available.[2] There exists a one-to-one correspondence between the samples of these two datasets.

For training purposes, we construct different types of training data using three strategies described in Section IV-C. For testing, we assume that there is an available English test set denoted as $\{(q_i^{en}, p_{i,1}^{en}, \ldots, p_{i,500}^{en})\}_{i=1}^{M}$ and its corresponding Chinese translation. For each English sample, we randomly replace, q_i^{en}, with its Chinese translation, q_i^{zh}, in a coin-tossing manner. Then we randomly select half of the passages (i.e., 250 passages) and replace them with their corresponding Chinese translations. Therefore, a new test sample is generated with the form of $(q_i^{en}, p_{i,1}^{en}, p_{i,2}^{zh}, \ldots, p_{i,499}^{en}, p_{i,500}^{zh})$ or $(q_i^{zh}, p_{i,1}^{en}, p_{i,2}^{zh}, \ldots, p_{i,499}^{en}, p_{i,500}^{zh})$.

3) ZERO-SHOT CROSS-LINGUAL PASSAGE RE-RANKING DATASETS

To further verify the robustness of the proposed approach, we directly transfer the trained model to test on three out-domain datasets that are created by ourselves. Specifically, since there is no publicly available dataset for XPR, we modify three available Question Answering (QA) datasets for this testing, including BiPaR,[3] MLQA,[4] and XQuAD.[5]

a: BiPaR

It is a bilingual parallel novel-style MRC dataset. BiPaR consists of 3,667 paragraphs and 14,668 question-answer pairs excerpted from Chinese and English novels. It has been divided into three parts: training (11,668 QA pairs), development (1,500 QA pairs), and test set (1,500 QA pairs).

b: MLQA

It is a multi-way aligned extractive QA evaluation benchmark. The dataset is constructed by mining parallel paragraphs from Wikipedia. It consists of question-answer pairs in seven languages, including English and Chinese. Especially, MLQA contains 5,641 extractive QA instances in Chinese and all instances are parallel with English. These instances have been divided into two parts: dev (504 QA instances) and test set (5,137 QA instances).

c: XQuAD

It is a cross-lingual QA dataset, which is translated from the development set of SQuAD v1.1. XQuAD consists of 240 paragraphs and 1,190 question-answer pairs in eleven languages, including English and Chinese. That is to say, each question appears in 11 different languages and has 11 parallel correct answers.

TABLE 2. Numbers of constructed queries and candidates for each query in InsuranceQA_v2, BiPaR, MLQA, and XQuAD.

Dataset	Queries	Candidates
InsuranceQA_v2	1625	500
BiPaR	1500	375
MLQA	504	443
XQuAD	1190	240

We modify these datasets using the same method of constructing insuranceQA_v2 test data. Notably, since the original test data from BiPaR and XQuAD have not been released, we instead use their development sets for out-domain testing. For each query within, the candidates to be ranked include all paragraphs across the corresponding dataset. If a passage contains the target answer, it is considered relevant to the given query. Table 2 shows the number of queries and candidates for each query we collect in BiPaR, MLQA, and XQuAD.

[2] https://github.com/Samurais/insuranceqa-corpus-zh

[3] https://github.com/sharejing/BiPaR

[4] https://github.com/facebookresearch/MLQA

[5] https://github.com/deepmind/xquad

TABLE 3. Results of applying different methods on cross-lingual passage re-ranking dataset.

#	Method	acc@1(%)	acc@10(%)	MRR(%)	MAP(%)
1	mBERT+Merged Training (baseline)	25.29	52.18	34.57	28.81
2	mBERT+Cascade Training	25.42	53.85	35.22	29.43
3	mBERT+Mixed Training	26.95	64.92	39.51	34.55
4	mBERT+XPI+Merged Training	26.83	59.14	37.72	32.14
5	mBERT+XPI+Cascade Training	24.49	56.80	35.34	29.91
6	mBERT+XPI+Mixed Training	**28.43**	**67.38**	**40.86**	**36.10**

B. EVALUATION METRICS

In this article, we adopt three metrics to evaluate the effectiveness of our proposed method, namely, top-k accuracy, Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). These metrics are widely used in the Information Retrieval (IR) context. In the following, we describe these metrics in detail.

1) TOP-K ACCURACY

Top-k accuracy, i.e. $acc@k$, measures the percentage of the queries with at least one relevant passage in the list, P_r, of top k ranked passages. Given M queries, $\{q_i\}_{i=1}^{M}$, $acc@k$ is obtained as:

$$acc@k = \frac{1}{M}\sum_{i=1}^{M} isExist(q_i), \tag{5}$$

and

$$isExist(q_i) = \begin{cases} 0 & \text{no relevant passage for } q_i \text{ in } P_r^{(i)} \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

where $P_r^{(i)}$ denotes the ranked list for the i^{th} query.

2) MAP

We first describe Precision which is defined as:

$$Precision@k(q_i) = \frac{\sum_{n=1}^{k} Rel(n)}{k}, \tag{7}$$

where $Rel(n)$ refers to the ground-truth relevance between the query, q_i, and the n^{th} passage, and k stands for the number of passages in the ranked list, $P_r^{(i)}$. The Average Precision (AP) [30] is then defined as:

$$AP(q_i) = \frac{\sum_{n=1}^{k} Presicion@n(q_i) \times Rel(n)}{\sum_{n=1}^{m} Rel(n)}, \tag{8}$$

where m refers to the number of the full list of passages for q_i. MAP is then defined as the mean of the average precision scores for each query across all queries:

$$MAP = \frac{1}{M}\sum_{i=1}^{M} AP(q_i). \tag{9}$$

3) MRR

Mean Reciprocal Rank [31] is also obtained using the binary relevance judgments and is defined as the reciprocal rank of the first relevant passage averaged across all queries. The Reciprocal Rank (RR) is defined as:

$$RR(q_i) = \frac{1}{rank_{q_i}}, \tag{10}$$

where $rank_{q_i}$ is the rank position of the first relevant passage for the i^{th} query. MRR is then defined as:

$$MRR = \frac{1}{M}\sum_{i=1}^{M} RR(q_i). \tag{11}$$

C. BASELINE

To prove the superiority of our method, we compare it with the pre-trained model mBERT[6]. mBERT is a multilingual version of BERT, which is trained on Wikipedia monolingual corpora in 104 languages. This model proves to be surprisingly effective in a wide range of cross-lingual tasks [32], [33], e.g., reading comprehension, document classification, etc.

For the baseline, we directly fine-tune mBERT using merged training as described in Section IV-C.

D. IMPLEMENTATION DETAILS

We adopt the base case mBERT as the basis for all experiments, which is a Transformer [26] with 12 layers, 12 heads, and GELU activation function. For the extended pre-training stage, we follow the original BERT implementation and train our model using Adam optimizer [34] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is set to 2e-5, and we further use the original linear decay. For the fine-tuning stage, we use the same optimizer and learning rate as in the extended pre-training. Each task is trained until convergence of the metric of the respective task. We then store the best model with the highest top-1 accuracy as well as the last model before terminating the training process.

VI. RESULTS AND ANALYSES

The results of the experiments and their analysis are presented in the following.

[6] https://github.com/google-research/bert

TABLE 4. Results of transferring the trained model on zero-shot cross-lingual passage re-ranking datasets.

Dataset	Method	acc@1(%)	acc@10(%)	MRR(%)	MAP(%)
BiPaR	mBERT+Mixed Training	15.13	39.50	23.41	23.41
	mBERT+XPI+Mixed Training	**17.07**	**42.73**	**25.95**	**25.95**
MLQA	mBERT+Mixed Training	29.96	58.93	39.48	39.48
	mBERT+XPI+Mixed Training	**33.33**	**64.68**	**43.04**	**43.04**
XQuAD	mBERT+Mixed Training	47.9	83.19	59.81	59.81
	mBERT+XPI+Mixed Training	**56.64**	**88.40**	**67.80**	**67.80**

A. RESULTS ON CROSS-LINGUAL PASSAGE RE-RANKING DATASET

The results on cross-lingual passage re-ranking test set are presented in Table 3. Amongst the training strategies, Mixed Training (#3) achieves the highest performance. Compared with the baseline (#1) in which we train mBERT using Merged Training, Mixed Training achieves 1.66%, 12.17%, 4.94%, 5.74% improvements in acc@1, acc@10, MRR and MAP, respectively. The most likely reason behind the archived gains is the higher similarity between the training data and the test data of the target task.

Regarding the combination of XPI with different training strategies, the combination of XPI and Mixed Training (#6) provides consistent improvements over Mixed Training. It outperforms the baseline by a large margin in all metrics and provides 3.14%, 15.20%, 6.29%, 7.29% improvements in acc@1, acc@10, MRR and MAP, respectively. These results confirm the capability of XPI in augmenting the cross-lingual alignment which is essential for our target task. This also suggests that continuing pre-training the model towards a specific task provides significant benefits, as also reported by Gururangan et.al [35]. Nevertheless, our results show that the combination of XPI and Cascade Training (#5) slightly reduces the model performance in terms of acc@1. A possible explanation for this is that XPI followed by Cascade Training might be prone to catastrophic forgetting of the cross-lingual alignment [36].

In summary, the above results show that alignment between languages, especially sentence-level alignment, is essential for XPR. Besides, the extra pre-training task, i.e., XPI, can effectively augment the cross-lingual alignment. It is also seen that exposure to a larger amount of training data which is similar to the test data, in the fine-tuning stage is beneficial for the performance of the targeted task.

B. RESULTS ON ZERO-SHOT CROSS-LINGUAL PASSAGE RE-RANKING DATASETS

To verify the cross-domain robustness of our method, we directly transfer the trained model that performs best on insuranceQA_v2 to test on three real-life-like datasets, including BiPaR, MLQA, and XQuAD. It is noteworthy that what we are measuring here is the model ability of passage re-ranking on out-domain data.

The performance of our method on three out-domain datasets is shown in Table 4. In agreement with our previous findings, the combination of XPI and Mixed Training provides significant improvements over Mixed Training in all metrics. These results once again support our claim that alignment between languages is essential for XPR, and also demonstrate that XPI can effectively augment the required alignment. Remarkably, the model performance on MLQA and XQuAD exceeds that achieved on BiPaR by a large margin. A possible explanation for this is that MLQA and XQuAD are collected from Wikipedia, while BiPaR is excerpted from novels. Thus, our model, which is built on top of mBERT pre-trained on a large-scale corpus of Wikipedia, performs better on these two datasets. This suggests that the more similar the domain of the training data, either for pretraining or for fine-tuning, is to that of the testing ones, the better performance can be achieved.

C. CASE STUDY

To demonstrate how the XPI affects the accuracy of XPR, here we conduct a case study as in Table 5. For brevity, we randomly choose three examples from the outputs of our model and provide the top three passages as instances for each query.

In Example 1, the question which is asked in English is: “What are the benefits of long term care insurance?” using mBERT without XPI, the relevant passage in Chinese “长期护理保险的一些好处是：保护您的遗产，选择满足您的长期护理需求...” is placed in the second position. However, mBERT with XPI can augment the cross-lingual alignment and help the relevant passage to be ranked in the first position. In Example 2, a similar situation is observed. The relevant passage which is placed in the third position by mBERT without XPI is ranked in the first position using XPI. The only difference is that the query and relevant passage are all in English. This suggests that although the candidate passages are in different languages, our model can find the correct one. In Example 3, XPI advantage in augmenting the cross-lingual alignment is indicated. As it is seen, using mBERT without XPI there is no relevant passage in the top three passages. Using XPI however, the relevant passage is ranked in the first position. This further suggests that XPI can narrow the gap in cases where the query and the passages to be ranked are from different languages. In summary, this case study confirms that our proposed method can augment the cross-lingual alignment and boost the model performance on the task of XPR.

TABLE 5. Case study. In each example, the passage in red color is the relevant one for the query.

Example 1	
Query:	What are the benefits of long term care insurance?
Top3 passages without XPI:	**No.1:** The cost of long term care insurance depends upon your gender, age, smoking status, health history... **No.2:** 长期护理保险的一些好处是：保护您的遗产，选择满足您的长期护理需求... (Some benefits of Long Term Care Insurance are: protection of your estate, having choices in meeting your long term care needs...) **No.3:** Long term care insurance premiums are determined by gender, age, smoking status, health history ...
Top3 passages with XPI:	**No.1:** 长期护理保险的一些好处是：保护您的遗产，选择满足您的长期护理需求... (Some benefits of Long Term Care Insurance are: protection of your estate, having choices in meeting your long term care needs...) **No.2:** 长期护理保险费是根据被保险人收集的医疗信息，如性别，年龄，健康史... (Long Term Care Insurance premium is determined by medical information collected on the insured like gender, age , heath history...) **No.3:** Long Term Care Insurance is an insurance policy that pays for the care you need if you are unable to care for yourself for extended periods of time...
Relevant passage:	长期护理保险的一些好处是：保护您的遗产，选择满足您的长期护理需求... (Some benefits of Long Term Care Insurance are: protection of your estate, having choices in meeting your long term care needs...)
Example 2	
Query:	Can a corporation pay for disability insurance?
Top3 passages without XPI:	**No.1:** 联邦政府为工作和支付系统的人提供社会残障保险... (The Federal Government provides Social Security Disability for people who worked and paid into the system...) **No.2:** 大多数大公司都有雇主集体福利计划，包括短期和长期残疾保险... (Most large corporations have employer group benefit plans that include short and long term disability insurance...) **No.3:** The business can pay for short or long term disability insurance in a employer sponsored group plan...
Top3 passages with XPI:	**No.1:** The business can pay for short or long term disability insurance in a employer sponsored group plan... **No.2:** 联邦政府为工作和支付系统的人提供社会残障保险... (The Federal Government provides Social Security Disability for people who worked and paid into the system...) **No.3:** Most disability plans will exclude preexisting conditions. A company is not going to pay for a disability caused by a known problem...
Relevant passage:	The business can pay for short or long term disability insurance in a employer sponsored group plan...
Example 3	
Query:	长期护理保险支付养老院费用吗? (Does long term care insurance pay for Nursing home?)
Top3 passages without XPI:	**No.1:** 长期护理保险是一种保险单，如果您无法长时间照顾自己，它支付您需要的护理费用... (Long Term Care Insurance is an insurance policy that pays for the care you need if you are unable to care for yourself for extended periods of time...) **No.2:** 如果这是唯一需要的护理，医疗保险不支付在养老院的护理费用... (Medicare does not pay for custodial care in a nursing home if that is the only care needed...) **No.3:** Long term care insurance is a relatively new form of insurance. It has been around since the early...
Top3 passages with XPI:	**No.1:** 是的，在某些情况下，长期护理保险可以支付养老院费用，或辅助生活，甚至在家庭服务中提供长期护理费用... (Yes, Long Term Care Insurance can pay for nursing home expenses, or for extended care offered by assisted living, or even in home services, in some situations...) **No.2:** Long Term Care Insurance policies will pay for as long as they are set up to pay... **No.3:** 私人医疗保险是否涵盖养老院? 不，不是真的。虽然可能会有一些情况，例如个人可能在疗养院... (Does private health insurance cover nursing homes? No, not really. While there may be situations such as transition where an individual may be at a nursing home...)
Relevant passage:	是的，在某些情况下，长期护理保险可以支付养老院费用，或辅助生活，甚至在家庭服务中提供长期护理费用... (Yes, Long Term Care Insurance can pay for nursing home expenses, or for extended care offered by assisted living, or even in home services, in some situations...)

VII. CONCLUSION AND DISCUSSION

In this article, we explored the Cross-lingual Passage Re-ranking (XPR) task which is designed to rank a list of candidate passages in multiple languages. We then propose Cross-lingual Paraphrase Identification (XPI) as an extra pre-training task that aims to further augment the cross-lingual alignment. Next, we modified a monolingual dataset with its translation to solve the shortage of large-scale annotated data. Furthermore, we presented three simple yet effective training strategies and directly transfer the trained model to test on three out-domain datasets. Experiments on the constructed datasets confirmed the effectiveness and robustness of our proposed method in conducting XPR. Note that the main focus of our article is on cross-lingual passage re-ranking between English and Chinese. In future works, we will incorporate extra languages into this task. Another research direction is to develop new pre-training tasks to help to augment the cross-lingual alignment and enhance the model performance, one step further.

REFERENCES

[1] C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power, "End-to-end neural ad-hoc ranking with kernel pooling," pp. 55–64, 2017.

[2] B. Mitra, F. Diaz, and N. Craswell, "Learning to match using local and distributed representations of text for Web search," pp. 1291–1299, 2017.

[3] K. Hui, A. Yates, K. Berberich, and G. De Melo, "Co-PACRR: A context-aware neural ir model for ad-hoc retrieval," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 279–287.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, vol. 1, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[5] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng, "A deep look into neural ranking models for information retrieval," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102067.

[6] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for Web search using clickthrough data," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2013, pp. 2333–2338.

[7] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 101–110.

[8] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2042–2050.

[9] Z. Dai, C. Xiong, J. Callan, and Z. Liu, "Convolutional neural networks for soft-matching N-Grams in ad-hoc search," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 126–134.

[10] R. Nogueira and K. Cho, "Passage re-ranking with BERT," 2019, *arXiv:1901.04085*. [Online]. Available: http://arxiv.org/abs/1901.04085

[11] W. Yang, H. Zhang, and J. Lin, "Simple applications of BERT for ad hoc document retrieval," 2019, *arXiv:1903.10972*. [Online]. Available: http://arxiv.org/abs/1903.10972

[12] D. A. Hull and G. Grefenstette, "Querying across languages: A dictionary-based approach to multilingual information retrieval," in *Proc. 19th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1996, pp. 49–57.

[13] G. Da San Martino, S. Romeo, A. Barroón-Cedeño, S. Joty, L. Maàrquez, A. Moschitti, and P. Nakov, "Cross-language question re-ranking," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 1145–1148.

[14] P. Zweigenbaum, S. Sharoff, and R. Rapp, "Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora," in *Proc. 10th Workshop Building Using Comparable Corpora*, 2017, pp. 60–67.

[15] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 597–610, Nov. 2019.

[16] U. Roy, N. Constant, R. Al-Rfou, A. Barua, A. Phillips, and Y. Yang, "LAReQA: Language-agnostic answer retrieval from a multilingual pool," 2020, *arXiv:2004.05484*. [Online]. Available: http://arxiv.org/abs/2004.05484

[17] A. Asai, J. Kasai, J. H. Clark, K. Lee, E. Choi, and H. Hajishirzi, "XOR QA: Cross-lingual open-retrieval question answering," 2020, *arXiv:2010.11856*. [Online]. Available: http://arxiv.org/abs/2010.11856

[18] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," 2013, *arXiv:1309.4168*. [Online]. Available: http://arxiv.org/abs/1309.4168

[19] S. Gouws, Y. Bengio, and G. Corrado, "Bilbowa: Fast bilingual distributed representations without word alignments," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 748–756.

[20] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," 2017, *arXiv:1710.04087*. [Online]. Available: http://arxiv.org/abs/1710.04087

[21] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7059–7069.

[22] A. Conneau, S. Wu, H. Li, L. Zettlemoyer, and V. Stoyanov, "Emerging cross-lingual structure in pretrained language models," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6022–6034.

[23] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.

[24] Y. Sun, S. Wang, Y.-K. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in *Proc. AAAI*, 2020, pp. 8968–8975.

[25] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4996–5001.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[27] J. Singh, B. McCann, N. Shirish Keskar, C. Xiong, and R. Socher, "XLDA: Cross-lingual data augmentation for natural language inference and question answering," 2019, *arXiv:1905.11471*. [Online]. Available: http://arxiv.org/abs/1905.11471

[28] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The united nations parallel corpus v1. 0," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, 2016, pp. 3530–3534.

[29] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou, "Applying deep learning to answer selection: A study and an open task," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 813–820.

[30] M. Zhu, "Recall, precision and average precision. Department of statistics and actuarial science, University of Waterloo," Waterloo Work. Paper, Tech. Rep., 2004.

[31] N. Craswell, "Mean reciprocal rank," in *Encyclopedia of Database Systems*, vol. 1703. 2009.

[32] T.-Y. Hsu, C.-L. Liu, and H.-Y. Lee, "Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5935–5942.

[33] S. Wu and M. Dredze, "Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 833–844.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[35] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 8342–8360. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.740

[36] D. Yogatama, C. de Masson d'Autume, J. Connor, T. Kocisky, M. Chrzanowski, L. Kong, A. Lazaridou, W. Ling, L. Yu, C. Dyer, and P. Blunsom, "Learning and evaluating general linguistic intelligence," 2019, *arXiv:1901.11373*. [Online]. Available: http://arxiv.org/abs/1901.11373

DONGMEI CHEN was born in Guangxi, China, in 1996. She received the bachelor's degree in information management and information system from Central South University (CSU), in 2018. She is currently pursuing the master's degree with the National University of Defense Technology (NUDT) and the master's degree in management science and engineering. Her research interests include natural language processing, data mining, and deep learning.

SHENG ZHANG received the B.S. degree in systems engineering and the M.S. degree in management science and engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2015 and2017, respectively, where he is currently pursuing the Ph.D. degree with the College of Systems Engineering. His research interests include natural language processing, deep learning, and data mining.

XIN ZHANG received the B.S. and Ph.D. degrees in system engineering from the National University of Defense Technology (NUDT), China, in 2000 and 2006, respectively, where he is currently a Professor with the Science and Technology on Information Systems Engineering Laboratory, College of Systems Engineering. His research interests include cross-model data mining, information extraction, and event analysis.

KAIJING YANG was born in Shandong, China, in 1996. She received the bachelor's degree in information and computing science from South-Central Minzu University (SCMZU), Wuhan, China, in 2018. She is currently pursuing the master's degree in management science and engineering with the National University of Defense Technology (NUDT). Her research interests include natural language processing, data mining, and text analysis.