

Cross-lingual Machine Reading Comprehension with Language Branch Knowledge Distillation

Junhao Liu^{1,2*} Linjun Shou^{3*} Jian Pei⁴ Ming Gong³ Min Yang^{1†} Daxin Jiang^{3†}

¹Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

²University of Chinese Academy of Sciences

³Microsoft STCA NLP Group, Beijing, China

⁴School of Computing Science, Simon Fraser University

{jh.liu,min.yang}@siat.ac.cn, {lisho,migon,djiang}@microsoft.com, jpei@cs.sfu.ca

Abstract

Cross-lingual Machine Reading Comprehension (CLMRC) remains a challenging problem due to the lack of large-scale annotated datasets in low-source languages, such as Arabic, Hindi, and Vietnamese. Many previous approaches use translation data by translating from a rich-source language, such as English, to low-source languages as auxiliary supervision. However, how to effectively leverage translation data and reduce the impact of noise introduced by translation remains onerous. In this paper, we tackle this challenge and enhance the cross-lingual transferring performance by a novel augmentation approach named Language Branch Machine Reading Comprehension (LBMRC). A language branch is a group of passages in one single language paired with questions in all target languages. We train multiple machine reading comprehension (MRC) models proficient in individual language based on LBMRC. Then, we devise a multilingual distillation approach to amalgamate knowledge from multiple language branch models to a single model for all target languages. Combining the LBMRC and multilingual distillation can be more robust to the data noises, therefore, improving the model’s cross-lingual ability. Meanwhile, the produced single multilingual model can apply to all target languages, which saves the cost of training, inference, and maintenance for multiple models. Extensive experiments on two CLMRC benchmarks clearly show the effectiveness of our proposed method.

1 Introduction

Machine Reading Comprehension (MRC) is a central task in natural language understanding (NLU) with many applications, such as information retrieval and dialogue generation. Given a query and a text paragraph, MRC extracts the span of the correct answer from the paragraph. Recently, as a series of large-scale annotated datasets become available, such as SQuAD (Rajpurkar et al., 2016) and TriviaQA (Joshi et al., 2017), the performance of MRC systems has been improved dramatically (Xiong et al., 2017; Hu et al., 2017; Yu et al., 2018; Wang et al., 2016; Seo et al., 2016). Nevertheless, those large-scale, high-quality annotated datasets often only exist in rich-resource languages, such as English, French and German. Correspondingly, the improvement of MRC quality can only benefit those rich-source languages. Annotating a large MRC dataset with high quality for every language is very costly and may even be infeasible (He et al., 2017). MRC in low-resource languages still suffers from the lack of large amounts of high-quality training data.

Besides, in real business scenarios, it is not practical to train separate MRC models for each language given that there are thousands of languages existed in the world. Thus multi-lingual MRC (single model for multiple languages) is of strong practical value by greatly reducing the model training, serving and maintenance costs.

To tackle the challenges of MRC in low-resource languages, cross-lingual MRC (CLMRC) is proposed, where translation systems are used to translate datasets from rich-source languages to enrich

*Equal contribution. Work was done when Junhao Liu was an intern at Microsoft STCA.

†Min Yang and Daxin Jiang are corresponding authors.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

training data for low-resource languages (Asai et al., 2018; Lee et al., 2018). However, CLMRC is severely restricted by translation quality (Cui et al., 2019).

Recently, large-scale pre-trained language models (PLM) (Devlin et al., 2018; Yang et al., 2019; Sun et al., 2019) are shown effective in NLU related tasks. Inspired by the success of PLM, multilingual PLM (Lample and Conneau, 2019; Huang et al., 2019; Liang et al., 2020) are developed by leveraging large-scale multilingual corpora for cross-lingual pre-training. Those powerful multilingual PLM are capable of zero-shot or few-shot learning (Conneau et al., 2018; Castellucci et al., 2019), and are effective to transfer from rich-resource languages to low-resource languages. Although those methods gain significant improvements on sentence-level tasks, such as sentence classification (Conneau et al., 2018), there is still a big gap between the performance of CLMRC in rich-resource languages and that in low-resource languages, since CLMRC requires high quality fine-grained representation at the phrase-level (Yuan et al., 2020).

Several studies combine multilingual PLM with translation data to improve the CLMRC performance by either data augmentation using translation (Singh et al., 2019) or auxiliary tasks (Yuan et al., 2020) (see Section 2 for some details). Those studies take two alternative approaches. First, they may just leverage translated data in target languages as new training data to directly train target language models (Hsu et al., 2019). The performance of such models is still limited by the translation issues (i.e., the noise introduced by the translation processing). Second, they may strongly rely on language-specific external corpora, which are not widely or easily accessible (Yuan et al., 2020).

According to the generalized cross-lingual transfer result (Lewis et al., 2019), the best cross-lingual performance is often constrained by the passage language, rather than the question language. In other words, the passage language plays an important role in CLMRC. The intuition is that the goal of MRC to pinpoint the exact answer boundary in passage, thus the language of passage has stronger influence on the performance than the question language. Motivated by this intuition, in this paper, we propose a new cross-lingual training approach based on knowledge distillation for CLMRC. We group the translated dataset (i.e., both questions and passages are translated into all target languages) into several groups. A group, called a *language branch*, contains all passages in one single language paired with questions in all target languages. For each language branch, a separate teacher model is trained. Those language branch specific models are taken as teacher models to jointly distill a single multilingual student model using a novel multilingual distillation framework. With this framework, our method can amalgamate multiple language diversity knowledge from language branch specific models to a single multilingual model and can be more robust to defeat the noises in the translated dataset, which obtains better cross-lingual performance.

We make the following technical contributions. First, on top of translation, we propose a novel language branch training approach by training several language specific models as teachers to provide fine-grained supervisions. Second, based on those teacher models, we propose a novel multilingual multi-teacher distillation framework to transfer the capabilities of the language teacher models to a unified CLMRC model. Last, we conduct extensive experiments on two popular CLMRC benchmark datasets in 9 languages under both translation and zero-shot conditions. Our model achieves state-of-the-art results on all languages for both datasets without using any external large-scale corpus.

The rest of the paper is organized as follows. We review related work in Section 2, and present our method in Section 3. We report experimental results in Section 4 and conclude the paper in Section 5.

2 Related Work

Our study is mostly related to the existing work on CLMRC and knowledge distillation. We briefly review some most related studies here.

Assuming only annotated data in another source language is available, CLMRC reads one context passage in a target language and extracts the span of an answer to a given question. Translation based approaches use a translation system to translate labeled data in a source language to a low-resource target language. Based on the translated data, Asai et al. (2018) devise a run-time neural machine translation based multilingual extracted question answering method. Singh et al. (2019) propose a data augmenta-

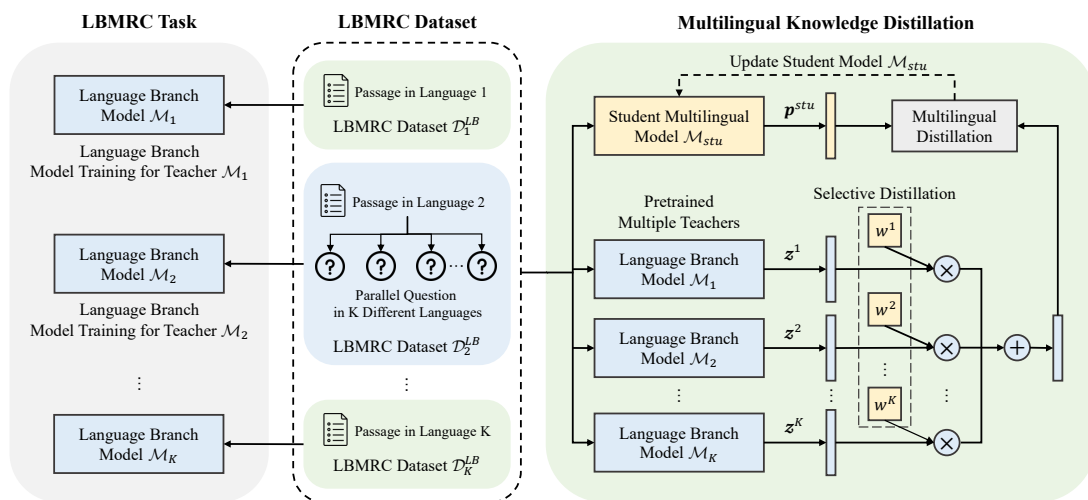


Figure 1: Model overview of our approach.

tion strategy for cross-lingual training. Cui et al. (2019) leverage a back-translation method to conduct CLMRC. All of these methods rely on a translation system to obtain high-quality translation data, which may not be available for some low-resource languages.

Recently, large-scale pre-trained language models have shown effective in many natural language processing tasks, which prompt the development of multilingual language models, such as multilingual BERT (Devlin et al., 2018), XLM (Lample and Conneau, 2019), and Unicoder (?). These language models aim to learn language agnostic contextual representations by leveraging large-scale monolingual and parallel corpuses, which show great potential on cross-lingual tasks, such as sentence classification tasks (Hsu et al., 2019; Pires et al., 2019; Conneau et al., 2018). However, there is still a big gap between the performance of CLMRC in rich-resource languages and that in low-resource languages, since CLMRC requires the capability of fine-grained representation at the phase-level (Yuan et al., 2020).

To further boost the performance of multilingual PLM on CLMRC task, Yuan et al. (2020) propose two auxiliary tasks mixMRC and LAKM on top of multilingual PLM. Those auxiliary tasks improve the answer boundary detection quality in low-resource languages. mixMRC first uses a translation system to translate the English training data into other languages and then constructs an augmented dataset of pairs $\langle \text{question}, \text{passage} \rangle$ in different languages. This new dataset turns out to be quite effective and can be used directly to train models on target languages. LAKM leverages language-specific meaningful phrases from external sources, such as entities mined from search logs of commercial search engines. LAKM conducts a new knowledge masking task. Any phrases contained in the training instances belonging to the external sources are replaced by a special token $[MASK]$. Then, the task of mask language model (Devlin et al., 2018) is conducted. The mixMRC task may still be limited by the translation quality and LAKM requires a large amount of external corpus, which is not easily accessible.

Knowledge Distillation is initially adopted for model compression (Buciluă et al., 2006), where a small and light-weight student model learns to mimic the output distribution of a large teacher model. Recently, knowledge distillation has been widely applied to many tasks, such as person re-identification (Wu et al., 2019), item recommendation (Tang and Wang, 2018), and neural machine translation (Tan et al., 2019; Zhou et al., 2019; Sun et al., 2020). Knowledge distillation from multiple teachers is also proposed (You et al., 2017; Yang et al., 2020), where the relative dissimilarity of feature maps generated from diverse teacher models can provide more appropriate guidance in student model training. Knowledge distillation is effective in transfer learning in those applications.

In this paper, on top of translation, we propose a novel approach of language branch training to obtain several language-specific teacher models. We further propose a novel multilingual multi-teacher distillation framework. In contrast to the previous work (Hu et al., 2018; Yuan et al., 2020), our proposed method can greatly reduce the noise introduced by translation systems without relying on external

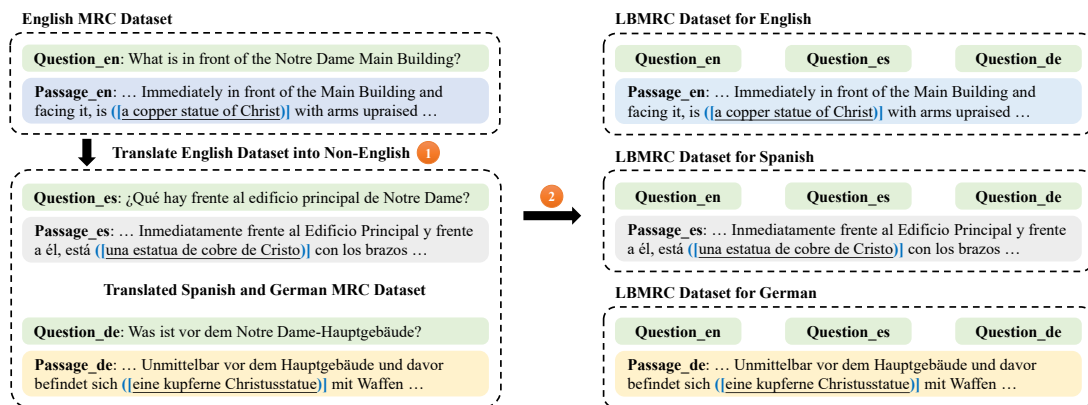


Figure 2: Overview of LBMRC dataset construction process. We use 3 languages (*English, Spanish, German*) in this illustration. In the first step, the English MRC dataset is translated into the other languages, including both questions and passages. In the second step, the construction method described in Section 3.1 is applied to build the LBMRC dataset for each language.

large-scale, language-specific corpus. Our method is applicable to more cross-lingual tasks.

3 Methodology

We formulate the CLMRC problem as follows. Given a labeled MRC dataset $\mathcal{D}_{src} = \{(p^{src}, q^{src}, a^{src})\}$ in a rich-resource language src , where p^{src} , q^{src} and a^{src} are a passage, a question, and an answer to q^{src} , respectively, the goal is to train a MRC model \mathcal{M} for the rich-resource language src and another low-resource language tgt . For an input passage p^{tgt} and question q^{tgt} in tgt , \mathcal{M} can predict the answer span $a^{tgt} = (a_s^{tgt}, a_e^{tgt})$, where a_s^{tgt} and a_e^{tgt} are the starting and ending indexes of the answer location in passage p^{tgt} , respectively. Model \mathcal{M} is expected to have good performance not only in the rich-resource language src , but also in the low-resource language tgt .

We first propose a new data augmentation based training strategy, Language Branch Machine Reading Comprehension (LBMRC), to train separate models for each language branch. A language branch is a group that contains passages in one single language accompanied with questions in all target languages. Under this setting, we can construct a language branch dataset for each language. Using each language branch dataset, we train a separate MRC model proficient in the language. Then, the branch-specific MRC models are taken as multiple teacher models to train a single multilingual MRC student model using a novel multilingual language branch knowledge distillation framework. The overview of our approach is illustrated in Figure 1.

3.1 Language Branch Machine Reading Comprehension (LBMRC)

The generalized cross-lingual transfer (G-XLT) approach (Lewis et al., 2019) trains a cross-lingual MRC model using the SQuAD (Rajpurkar et al., 2016) dataset and evaluates the model on samples of questions and passages in different languages. The results show that the best cross-lingual answering performance in the testing phase is sensitive to the language of passages in the test data rather than the language of questions. This observation suggests that the language of passages in training data may play an important role in the CLMRC task.

Based on the above understanding, we devise a new data augmentation based training strategy LBMRC. It first trains MRC models in several languages and then distills those models to derive a final MRC model for all target languages. In contrast to the mixMRC strategy (Yuan et al., 2020), LBMRC groups the translation data into several language branches using passage languages as identifiers. Each language branch contains all passages translated into one single language accompanied with questions in different languages. Figure 2 shows the overall procedure of this data construction process. We train a separate MRC model for each language branch, which is expected to be proficient in one specific language.

3.1.1 Language Branch Construction

To obtain parallel question and passage pairs in different languages, we adopt a method similar to (Yuan et al., 2020; Singh et al., 2019) by employing a machine translation system to translate a labeled dataset of questions and passages in English into datasets in multiple languages $\mathcal{D}_k = \{(p^k, q^k, a^k)\}$, where p^k, q^k and a^k are a passage, a question, and the answer to q^k , respectively, all in language k . In this process, it is hard to recover the correct answer spans in translated passages. To mitigate this problem, we take a method similar to (Lee et al., 2018) that adds a pair of special tokens to denote the correct answer in the original passage. We discard those samples where the answer spans cannot be recovered. The language branch for language k is the set of passages and answers in language k accompanied by the queries in all languages, that is, $\mathcal{D}_k^{LB} = \{(p^k, \{q^1, \dots, q^K\}, a^k)\}$, where K is the total number of languages.

3.1.2 Language Branch Model Training

Similar to the MRC training method proposed in BERT (Devlin et al., 2018), the PLM model is adopted for encoding the input text $x = [q, p]$ into a deep contextualized representation $\mathbf{H} \in \mathbb{R}^{L \times h}$, where L represents the length of the input text x , h is the hidden size of the PLM model. Then, we can calculate the final start and end position predictions $\mathbf{p}_s, \mathbf{p}_e$. Take the start position \mathbf{p}_s as an example, it can be obtained by the following equations:

$$\begin{aligned} \mathbf{z}_s &= \mathbf{H} \cdot \mathbf{u}_s + \mathbf{b}_s \\ \mathbf{p}_s &= \text{softmax}(\mathbf{z}_s / \tau) \end{aligned} \quad (1)$$

where $\mathbf{u}_s \in \mathbb{R}^h, \mathbf{b}_s \in \mathbb{R}^L$ are two trainable parameters, $\mathbf{z}_s \in \mathbb{R}^L$ represents the output logits, $\mathbf{p}_s \in \mathbb{R}^L$ is the predicted output distribution of the start positions, τ is the temperature introduced by (Hinton et al., 2015) to control the smoothness of the output distribution.

For each \mathcal{D}_k^{LB} , we train a language branch MRC model \mathcal{M}_k by optimizing the log-likelihood loss function:

$$\mathcal{L}_{NLL}(\mathcal{D}_k^{LB}; \mathcal{M}_k) = -\frac{1}{N} \sum_{i=1}^N \left((\mathbf{a}_{s,i}^k)^T \cdot \log(\mathbf{p}_{s,i}^k) + (\mathbf{a}_{e,i}^k)^T \cdot \log(\mathbf{p}_{e,i}^k) \right) \quad (2)$$

where N is the total number of samples in \mathcal{D}_k^{LB} , the temperature parameter τ is set to 1, $(\mathbf{p}_{s,i}^k, \mathbf{p}_{e,i}^k) \in \mathbb{R}^L$ are the start and end position predictions of sample i from model \mathcal{M}_k , $(\mathbf{a}_{s,i}^k, \mathbf{a}_{e,i}^k) \in \mathbb{R}^L$ are the ground-truth one-hot labels for the start and end positions of sample i in \mathcal{D}_k^{LB} .

3.2 Multilingual Multi-teacher Distillation

Let \mathcal{M}_{stu} denote the model parameters of the student multilingual MRC model. \mathcal{M}_{stu} is expected to distill the language-specific knowledge from the multiple language branch teachers $\{\mathcal{M}_k\}_{k=1}^K$. In terms of training data, we take the union of LBMRC datasets as the distillation training dataset \mathcal{D} which is $\mathcal{D} = \{\mathcal{D}_1^{LB} \cup \dots \cup \mathcal{D}_K^{LB}\}$.

Distillation Training We train a multilingual student model to mimic the output distribution of the language branch teacher models. Specifically, the distillation loss of the student model can be described in the form of cross-entropy. In order to distill knowledge from multiple teachers simultaneously, we propose to aggregate the predicted logits from different teachers. Formally, the distillation soft logits $\mathbf{z}_{s,i}, \mathbf{z}_{e,i}$ used to train the student model can be formulated as:

$$\mathbf{z}_{s,i} = \sum_{k=1}^K w_s^k \cdot \mathbf{z}_{s,i}^k; \quad \mathbf{z}_{e,i} = \sum_{k=1}^K w_e^k \cdot \mathbf{z}_{e,i}^k \quad (3)$$

where $w^k = \{w_s^k, w_e^k\}$ are hyper parameters to control the contributions of each teacher model, $\mathbf{z}_{s,i}^k$ and $\mathbf{z}_{e,i}^k$ are the predicted soft logits of sample i from the language branch teacher model \mathcal{M}_k . The

multilingual multi-teacher distillation loss can be calculated as:

$$\mathcal{L}_{KD}(\mathcal{D}; \mathcal{M}_{stu}) = -\frac{1}{N} \sum_{i=1}^N \left((\mathbf{p}_{s,i})^T \cdot \log(\mathbf{p}_{s,i}^{stu}) + (\mathbf{p}_{e,i})^T \cdot \log(\mathbf{p}_{e,i}^{stu}) \right) \times \tau^2 \quad (4)$$

where τ is the temperature parameter, $\mathbf{p}_{s,i}$ and $\mathbf{p}_{e,i}$ are the start and end distributions calculated by Equation 1 based on soft logits $\mathbf{z}_{s,i}, \mathbf{z}_{e,i}$, $\mathbf{p}_{s,i}^{stu}$ and $\mathbf{p}_{e,i}^{stu}$ are also calculated by the softmax-temperature based on the student predicted soft logits $\mathbf{z}_{s,i}^{stu}, \mathbf{z}_{e,i}^{stu}$.

Besides, the student model can be also trained using the ground-truth labels of start and end indexes. Let $\mathcal{L}_{NLL}(\mathcal{D}; \mathcal{M}_{stu})$ denote the log-likelihood loss function of the one-hot label on the training dataset \mathcal{D} , which can be formulated as follows:

$$\mathcal{L}_{NLL}(\mathcal{D}; \mathcal{M}_{stu}) = -\frac{1}{N} \sum_{i=1}^N \left((\mathbf{a}_{s,i})^T \cdot \log(\mathbf{p}_{s,i}^{stu}) + (\mathbf{a}_{e,i})^T \cdot \log(\mathbf{p}_{e,i}^{stu}) \right). \quad (5)$$

Finally, the whole multilingual distillation training loss for the student model \mathcal{M}_{stu} can be summarized as:

$$\mathcal{L}_{Total}(\mathcal{D}; \mathcal{M}_{stu}) = \lambda_1 \mathcal{L}_{NLL}(\mathcal{D}; \mathcal{M}_{stu}) + \lambda_2 \mathcal{L}_{KD}(\mathcal{D}; \mathcal{M}_{stu}) \quad (6)$$

where λ_1 and λ_2 are hyper parameters to balance the contribution of two types of loss.

Selective Distillation Here, we consider a proper mechanism to choose the distillation weights $\{w^k\}_{k=1}^K$ which can assist the student model to learn from a suitable teacher. We investigate two selection strategies and experiment with their performance in the distillation processing. As the first method, we treat the weights as prior hyper parameters which means that we fix the $\{w^k\}_{k=1}^K$ with initial values and train our student model with the same weights during the whole process. In the second mechanism, we use the entropy impurity to measure the teacher’s confidence of a predicted answer including the output distributions of start and end indexes. The confidence of the answer is higher when the impurity has a lower value. Take the start position aggregation as an example, the impurity value is used to determine the weight distribution $\{w_s^k\}_{k=1}^K$ as follows:

$$\begin{aligned} \mathbf{p}_s^k &= \text{softmax}(\mathbf{z}_s^k), \\ \mathcal{I}(\mathbf{p}_s^k) &= -(\mathbf{p}_s^k)^T \cdot \log(\mathbf{p}_s^k) \\ w_s^k &= \frac{\exp(\mathcal{I}(\mathbf{p}_s^k))}{\sum_j \exp(\mathcal{I}(\mathbf{p}_s^j))} \end{aligned} \quad (7)$$

where $\mathcal{I}(\cdot)$ represents the impurity function, \mathbf{z}_s^k represents the predicted logits of start position from \mathcal{M}_k . Based on this, the distillation weights for each teacher model can be adjusted automatically for each instance.

4 Experiments

Extensive experiments of our proposed method are conducted on two public cross-lingual MRC datasets. In the following sections, we describe our experimental settings, results, and analyze the performance.

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Metrics

To verify the effectiveness of our method. We use the following datasets to conduct our experiments.

MLQA A cross-lingual machine reading comprehension benchmark (Lewis et al., 2019). The instances in MLQA cover 7 languages. We evaluate our method on three languages (*English, German, Spanish*) with translation training method, and also test our method under the setting of zero-shot transfer on the other three languages (*Arabic, Hindi, Vietnamese*).

XQuAD Another cross-lingual question answering dataset (Artetxe et al., 2019). XQuAD contains instances in 11 languages, and we cover 9 languages in our experiments. Similar to the setting above, we evaluate our method on *English, German, Spanish*. In addition, we test our method on *Arabic, Hindi, Vietnamese, Greek, Russian, and Turkish* under the setting of zero-shot transfer.

Evaluation Metrics The evaluation metrics used in our experiments are same as the SQuAD dataset (Rajpurkar et al., 2016) including F1 and Exact Match score. F1 score measures the answer overlap between the predicted and ground-truth answer spans. Exact Match score measures the percentage of predicted answer spans exactly matching the ground-truth labels. We use the official evaluation script provided by (Lewis et al., 2019) to measure performance over different languages. For the XQuAD dataset, we follow the official instruction provided by (Artetxe et al., 2019) to evaluate our predicted result.

4.1.2 Baseline Methods

We compare our method with the following baseline methods: (1) Baseline, a method originally proposed in (Lewis et al., 2019) that the MRC model is trained in *English* dataset and tested on the other languages directly, (2) mixMRC, a translation based data augmentation strategy proposed in (Yuan et al., 2020; Singh et al., 2019), which mixes the question and passage in different languages, (3) LAKM, a pre-trained task devised in (Yuan et al., 2020) by introducing external sources for phrase level mask language model task, and (4) mixMRC + LAKM, a combination method of (2) and (3) through multiple task learning.

4.1.3 Implementation Details

We adopt the pre-trained multilingual language model XLM (Lample and Conneau, 2019) to conduct our experiments. XLM is a cross-lingual language model pre-trained with monolingual and parallel cross-lingual data to achieve decent transfer performance on cross-lingual tasks. We use the Transformers library from HuggingFace (Wolf et al., 2019) to conduct our experiments. For the MRC task, the pre-trained model is used as the backbone and two trainable vectors are added to locate the start and end positions in the context passage, same with (Devlin et al., 2018).

To construct the LBMRC dataset, We translate the SQuAD dataset to *Spanish* and *German* languages which are two relatively high-resource languages, hence, the number of language branch models is 3 ($K = 3$). The target languages of our CLMRC model are *English, Spanish, and German*. The English branch dataset is always added to other non-English language branch datasets to improve the data quality, which can reduce the impact of noise in data and improve the performance of non-English teachers in our experiments.

In order to fit the multilingual model into the GPU memory, we pre-processed the teachers’ logits for each instance in dataset \mathcal{D} . For the multilingual model training, We use AdamW optimizer with $\text{eps} = 1e^{-8}$ and set weight decay to 0.005. The learning rate is set as $1e^{-5}$ for the language branch model training and distillation training. The XLM model is configured with its default setting. For the first selective distillation mechanism, we set the hyper parameters of $w_s^k = w_e^k = 1/K$ which reach the best performance in our experiments. The distillation loss weight is set as $\lambda_1 = 0.5, \lambda_2 = 0.5$ and the softmax temperature $\tau = 2$. We train 10 epochs for each task which can make sure each task converges.

4.2 Experimental Results

4.2.1 Results on MLQA

We first evaluate our method on the MLQA dataset in 6 languages. The results are shown in Table 1. Compared with XLM baselines of original report results and our reproduced results, our method with both selective multilingual distillation strategies (Ours-hyper, Ours-imp) outperform the strong baseline LAKM, mixMRC and mixMRC + LAKM in en, es and de. Especially note that the LAKM method uses extra language corpus to train a better backbone language model, while our method without using external data can also improve the performance significantly with more than 3% consistent gains in es and de languages. This verifies that the LBMRC training approach could preserve the language characteristics in each teacher model and the multi-teacher distillation step could further reduce the training data noise introduced during the translation process.

Methods	MLQA (EM / F1)					
	en	es	de	ar	hi	vi
Lewis [¶]	62.4 / 74.9	47.8 / 65.2	46.7 / 61.4	34.4 / 54.0	33.4 / 50.7	39.4 / 59.3
Baseline	63.4 / 77.3	49.7 / 68.2	48.5 / 63.7	37.5 / 56.9	37.0 / 54.3	42.4 / 63.5
LAKM	64.6 / 79.0	52.2 / 70.2	50.6 / 65.4	-	-	-
mixMRC	63.8 / 78.0	52.1 / 69.9	49.8 / 64.8	38.5 / 58.4	40.1 / 57.1	45.2 / 66.2
mixMRC + LAKM	64.4 / 79.1	52.2 / 70.3	51.2 / 66.0	-	-	-
Ours-hyper	64.8 / 79.3	53.9 / 71.8	52.1 / 66.8	40.4 / 60.0	42.8 / 59.8	46.1 / 67.2
Ours-imp	64.7 / 79.2	54.3 / 72.0	52.4 / 66.9	40.1 / 59.9	42.9 / 59.9	46.5 / 67.5

Table 1: EM and F1 score of 6 languages on the MLQA dataset. The left 3 languages (en, es, de) are under translation condition while the right part (ar, hi, vi) results are under the zero-shot transfer method. The results with [¶] are adopted from Lewis et al. (2019).

We further test our method under the setting of zero-shot transfer in other languages ar, hi, vi. Since the LAKM method requires language-specific corpora to train the backbone model, it is not feasible to access such a corpus to train the backbone model for every low-resource language. Hence, we only compare our method with mixMRC for a fair comparison. For the languages ar, hi and vi, we zero-shot transfer our model to predict in these contexts. We can find that our method also obtains state-of-art results compared with the mixMRC and Baseline with more than 4% improvement. These results in the MLQA dataset show that our method not only improves the performance in the languages included in our language branch training but also has better-transferring capability to predict the answer in those languages not included in our language branch.

To compare the two selective distillation strategies we devised above, the impurity selective mechanism Ours-imp gets the best results on most languages, thus proving to be a proper way to aggregate the knowledge from multiple language branch teachers than the weight fixing method Ours-hyper.

Methods	XQuAD (EM / F1)		
	en	es	de
Baseline	68.8 / 81.3	56.9 / 75.6	55.5 / 72.6
mixMRC	69.2 / 82.4	58.7 / 78.8	58.2 / 75.4
Ours-hyper	69.9 / 83.2	59.3 / 79.9	59.4 / 76.3
Ours-imp	70.1 / 83.4	59.6 / 80.0	59.8 / 76.5

Table 2: EM and F1 score of 3 languages on the XQuAD dataset under the translation-train setting.

Methods	XQuAD (EM / F1)					
	ar	hi	vi	el	ru	tr
Baseline	43.2 / 62.6	46.0 / 63.1	48.7 / 70.4	49.4 / 68.5	55.2 / 72.3	44.1 / 63.1
mixMRC	42.4 / 63.6	50.0 / 66.2	52.7 / 72.6	51.1 / 72.1	58.7 / 75.9	47.8 / 65.8
Ours-hyper	44.5 / 65.0	52.0 / 67.4	55.5 / 74.6	52.2 / 73.1	59.3 / 76.6	50.8 / 68.3
Ours-imp	44.0 / 64.6	52.5 / 67.9	55.6 / 74.9	52.4 / 73.3	59.6 / 76.6	50.2 / 67.7

Table 3: EM and F1 score of 6 languages on the XQuAD dataset under the zero-shot transfer setting.

4.2.2 Results on XQuAD

We evaluate our method on another common used cross-lingual benchmark XQuAD dataset in 9 languages. The results are shown in Table 2 and 3 which are under the condition of translation and zero-shot respectively. Since the LAKM method is not suitable in this dataset, we directly compare our method

with the mixMRC. Our method consistently outperforms the mixMRC methods in both two conditions. In terms of translation condition, our best method Ours-imp gets 1.7% and 2.7% improvement of EM score on es and de respectively. The impurity selective strategy is better for these 3 languages. In terms of the zero-shot transfer, our method obtains a bigger improvement in these 6 languages. Take the Ours-hyper as an example, 4 languages (ar, hi, vi, tr) gain more than 2 points increase of EM score compared with the strong baseline mixMRC. The other 2 languages also have decent EM metric improvement with 2.5% and 1.5% for el and ru respectively. The evaluation results on the XQuAD dataset further verify the effectiveness and robustness of our proposed method.

4.3 Analysis

4.3.1 Why Use LBMRC?

First, we analyze the training strategy of LBMRC. In order to get different language branch teachers, a possible way is to train each teacher model according to the translated dataset directly. Take the German as an example, this method directly uses the German MRC dataset translated from English as the training set to train the German MRC model. The evaluation results are shown in Table 4, where the results with ¶ are adopted from Lewis et al. (2019), the Baseline method uses the only *English* dataset to train and then zero-shot transfers to other languages, and the w/o LBMRC method translates the *English* dataset to other languages and train separate MRC models for each translated dataset. This method (w/o LBMRC) even underperforms the Baseline method where only English training data is used, which aligns with the results from Lewis et al. (2019). The results show that simply leveraging the noisy translated datasets for training can not improve the performance properly.

Methods	MLQA (EM / F1)			Teachers	MLQA (EM / F1)		
	en	es	de		en	es	de
Lewis¶	62.4 / 74.9	49.8 / 68.0	47.6 / 62.2	En	63.2 / 77.6	51.2 / 69.4	49.3 / 64.2
Baseline	63.4 / 77.3	49.7 / 68.2	48.5 / 63.7	Es	62.9 / 77.4	51.9 / 70.0	49.1 / 63.9
w/o LBMRC	-	49.6 / 68.3	47.5 / 62.2	De	63.6 / 78.1	51.2 / 69.3	50.4 / 65.0

Table 4: Translate-train result instead of LBMRC. Table 5: Performance of LBMRC teacher models.

The performance of our LBMRC teacher models are shown in Table 5. With the method introduced in Section 3.1.2, we train each language branch teacher model using the according LBMRC dataset. From the results, we can see that the Es and De teacher models achieve the best result on the test set in its own language, which verifies the hypothesis we proposed in Section 3.1.2 that teacher model trained using LBMRC can preserve language-specific characteristics. Compared with models trained using the mixMRC strategy, LBMRC preserve the language diversity to obtain language-specific expert models.

4.3.2 Why Multilingual Multi-teacher Distillation Works?

According to Table 5, an observation is the performance of our teachers is worse than our distillation models (Ours-hyper, Ours-imp) in Table 1, which due to the hidden noise in the training set introduced by the translation process. With the help of multilingual distillation training, the student model can be more robust to the data noises and effective to use the translated dataset. We further conduct some ablation studies on different teacher settings: (1) w/o de, remove the de teacher model during multilingual distillation training process, (2) w/o es, remove the es teacher model during multilingual distillation, (3) w/ en, only adopt the en teacher model into the multilingual distillation process, and (4) w/ mix, we take three MRC models trained with mixMRC strategy as the teacher models to do distillation training and obtain a new single student model, where we use the same number of the teacher in our method for a fair comparison. This study (w/ mix) is to verify the effectiveness of the language branch-based multilingual distillation. The ablation results are reported in Table 6.

With the ablation study results, we can summarize that each teacher in different languages can have specific contributions to our approach. Take w/o de as an example, the result shows that the de result drops significantly compared with our best score while the en and es results are still relatively close. While the w/o es shows similar trends in terms of the es test result. For w/ en (without leveraging the

Exp.	MLQA (EM / F1)		
	en	es	de
Ours	64.7 / 79.2	54.3 / 72.0	52.4 / 66.9
- w/o de	64.5 / 79.0	54.1 / 71.9	51.2 / 66.1
- w/o es	64.4 / 78.6	53.2 / 71.2	51.8 / 66.6
- w/ en	63.9 / 78.2	52.8 / 70.9	50.9 / 65.8
- w/ mix	64.6 / 79.1	53.6 / 71.3	50.8 / 65.5

Table 6: The ablation study under different experiment settings.

knowledge from language branch teachers), the results degrade significantly on all languages. To further verify the importance of LBMRC in the multilingual distillation, we replace LBMRC teacher models with models trained using the mixMRC method. The experiment (w/ mix vs Ours) shows that the student model has similar performance in en, but the performance in es and de have a big gap compared with our method, especially for de. This shows that LBMRC could enhance cross-lingual transfer capability and the effectiveness of multilingual distillation.

5 Conclusions

In this paper, we propose a novel language branch data augmentation based training strategy (LBMRC) and a novel multilingual multi-teacher distillation framework to boost the performance of cross-lingual MRC in low-resource languages. Extensive experiments on two multilingual MRC benchmarks verify the effectiveness of our proposed method either in translation or zero-shot settings. We further analyze the reason why combine the LBMRC and multilingual distillation can gain better cross-lingual performance, which shows that our method is more effective to use the translation dataset and more robust to the noise hidden in the translated data. In addition, our distillation framework produces a single multilingual model applicable to all target languages, which is more practical to deploy multilingual serves.

Acknowledgments

Jian Pei’s research is supported in part by the NSERC Discovery Grant program. Min Yang’s research is partially supported by National Natural Science Foundation of China (No. 61906185), the Natural Science Foundation of Guangdong Province of China (No. 2019A1515011705), Shenzhen Basic Research Foundation (No. JCYJ20180302145607677), the Youth Innovation Promotion Association of CAS. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.
- Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275*.
- Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.
- Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multi-lingual intent detection and slot filling in a joint bert-based model. *arXiv preprint arXiv:1907.02884*.
- Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. *arXiv preprint arXiv:1909.00361*.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Tsung-Yuan Hsu, Chi-liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. *arXiv preprint arXiv:1909.09587*.
- Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2017. Reinforced mnemonic reader for machine reading comprehension. *arXiv preprint arXiv:1705.02798*.
- Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. 2018. Attention-guided answer distillation for machine reading comprehension. *arXiv preprint arXiv:1808.07644*.
- H. Huang, Yaobo Liang, N. Duan, Ming Gong, Linjun Shou, D. Jiang, and M. Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *EMNLP/IJCNLP*.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
- Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. 2018. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.
- Yaobo Liang, N. Duan, Yeyun Gong, N. Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, G. Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, S. Liu, Fan Yang, Rangan Majumder, and M. Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *ArXiv*, abs/2004.01401.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.
- Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.
- Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. *arXiv preprint arXiv:2004.10171*.
- Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*.
- Jiaxi Tang and Ke Wang. 2018. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2289–2298.

- Zhiguo Wang, Haitao Mi, Wael Hamza, and Radu Florian. 2016. Multi-perspective context matching for machine comprehension. *arXiv preprint arXiv:1612.04211*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.
- Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. 2019. Distilled person re-identification: Towards a more scalable system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1187–1196.
- Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.
- Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2020. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. *Proceedings of the 13th International Conference on Web Search and Data Mining*.
- Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294.
- Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.
- Fei Yuan, Linjun Shou, X. Bai, Ming Gong, Y. Liang, N. Duan, Y. Fu, and Daxin Jiang. 2020. Enhancing answer boundary detection for multilingual machine reading comprehension. In *ACL*.
- Chunting Zhou, Graham Neubig, and Jiatao Gu. 2019. Understanding knowledge distillation in non-autoregressive machine translation. *arXiv preprint arXiv:1911.02727*.