Sharing, Teaching and Aligning: Knowledgeable Transfer Learning for Cross-Lingual Machine Reading Comprehension

Tingfeng Cao¹²³, Chengyu Wang², Chuanqi Tan², Jun Huang², Jinhui Zhu¹³
¹School of Software Engineering, South China University of Technology, China
²Alibaba Group, China
³Key Laboratory of Big Data and Intelligent Robot (South China University of Technology) Ministry of Education, China
setingfengcao@mail.scut.edu.cn, csjhzhu@scut.edu.cn
{chengyu.wcy, chuanqi.tcq, huangjun.hj}@alibaba-inc.com

Abstract

In cross-lingual language understanding, machine translation is often utilized to enhance the transferability of models across languages, either by translating the training data from the source language to the target, or from the target to the source to aid inference. However, in cross-lingual machine reading comprehension (MRC), it is difficult to perform a deep level of assistance to enhance cross-lingual transfer because of the variation of answer span positions in different languages. In this paper, we propose X-STA, a new approach for cross-lingual MRC. Specifically, we leverage an attentive teacher to subtly transfer the answer spans of the source language to the answer output space of the target. A Gradient-Disentangled Knowledge Sharing technique is proposed as an improved cross-attention block. In addition, we force the model to learn semantic alignments from multiple granularities and calibrate the model outputs with teacher guidance to enhance cross-lingual transferability. Experiments on three multi-lingual MRC datasets show the effectiveness of our method, outperforming state-of-the-art approaches.

1 Introduction

Recently, significant progress has been made in NLP by pre-trained language models (PLMs) (Radford et al., 2018; Devlin et al., 2019; Zhang et al., 2022). Yet, these models often require a sufficient amount of training data to perform well, which is difficult to achieve in cross-lingual low-resource adaptation. Although many cross-lingual PLMs have been proposed to learn generic feature representations (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021; Liu et al., 2020), the performance gap between source and target languages is still relatively large, especially for token-level tasks such as machine reading comprehension (MRC). In addition, ultra-large PLMs such as ChatGPT (OpenAI, 2023) exhibit amazing zero-shot generation abilities over multiple languages. We observe that such models may not be sufficient for cross-language MRC due to the linguistic and cultural differences between these languages, together with the requirements of very fine-grained extraction of answer spans.

One of the most significant challenges in cross-lingual MRC is the lack of annotated datasets in low-resource languages, which are difficult to obtain. As seen, most of the current MRC datasets are in English (Rajpurkar et al., 2016). Another
challenge is the linguistic and cultural variations that exist across different languages, which exhibit different sentence structures, word orders and morphological features. For instance, languages such as Japanese, Chinese, Hindi and Arabic have different writing systems and a more complicated grammatical system than English, making it challenging for MRC models to comprehend the texts.

In the literature, machine translation-based data augmentation is often employed to translate the dataset of the source language into each target language for model training (Conneau et al., 2018; Hu et al., 2020; Ruder et al., 2021). As shown in Figure 1, it is relatively easy to enhance cross-lingual transferability of simple sequential classification tasks by directly fitting the output probability distribution of the source language via Kullback-Leibler Divergence (Fang et al., 2021; Zheng et al., 2021; Yang et al., 2022). However, for MRC, it is not possible to use the output distribution of the source language directly to teach the target language, due to the answer span shift caused by translation.

Motivated by this, we propose X-STA, a new approach for cross-lingual MRC that follows three principles: Sharing, Teaching and Aligning. For sharing, we propose the Gradient-Disentangled Knowledge Sharing (GDKS) technique, which uses parallel language pairs as model inputs and extracts knowledge from the source language. It enhances the understanding of the target language while avoiding degradation of the source language representations. For teaching, our approach leverages an attention mechanism by finding answers span from the target language’s context that are semantically similar to the source language’s output answers to calibrate the output answers. For aligning, alignments at multi-granularity are utilized to further enhance the cross-lingual transferability of the MRC model. In this way, we can enhance the language understanding of the model for different languages through knowledge sharing, teacher-guided calibration and multi-granularity alignment.

In summary, the main contributions of this study are as follows:

• We propose X-STA, a new approach for cross-lingual MRC based on three principles: sharing, teaching, and aligning.

• In X-STA, a Gradient-Disentangled Knowledge Sharing technique is proposed for transferring language representations. Output calibration and semantic alignments are further leveraged to enhance the cross-lingual transferability of the model.

• Extensive experiments on three multi-lingual MRC datasets verify that our approach outperforms state-of-the-art methods. Thorough ablation studies are conducted to understand the impact of each component of our method.

2 Related Work

In this section, we summarize the related work in the following three aspects.

2.1 Pre-trained Multi-lingual Language Models

Recent work has demonstrated that large-scale PLMs have tremendous potential for downstream tasks, as well as for multilingual representations including multilingual BERT (mBERT, Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-RoBERTa (Conneau et al., 2020), mT5 (Xue et al., 2021), mBART (Liu et al., 2020). These models extend training sets to unlabeled multilingual corpora and project all languages into the same semantic space, allowing for cross-lingual understanding.

2.2 Cross-lingual Knowledge Transfer

It aims to transfer knowledge learned from a source language to target languages. A intuitive approach is to use machine translation for data augmentation (Conneau et al., 2018; Bornea et al., 2021; Hu et al., 2020). Under this setting, more transferable cross-lingual representations can be learned through feature fusion (Fang et al., 2021), consistency regularization (Zheng et al., 2021) and manifold mixup (Yang et al., 2022). However, these works are not sufficiently exploited on translation data for MRC. Other work learns language-agnostic representations through adversarial training to explicitly decompose language-specific representations (Keung et al., 2019; Chen et al., 2019; Wu et al., 2022a), or through normalization to implicitly preserve more generic representations across languages (Libovický et al., 2020; Zhao et al., 2021; Aboagye et al., 2022). A more intuitive idea used for alignment is contrastive learning (Li et al., 2021; Feng et al., 2022; Zhang et al., 2023), where translation pairs are positive examples and texts from other pairs as negative examples.
2.3 Cross-lingual MRC

Yuan et al. (2020) propose several auxiliary pre-training tasks for solving answer boundary problems for low-resource languages. Liang et al. (2021) introduce an unsupervised phrase boundary recovery pre-training task to further address this problem. Chen et al. (2022) propose a two-stage step-by-step algorithm for finding the best answer from good to best for cross-lingual MRC. Wu et al. (2022b) introduce a Siamese Semantic Disentanglement Model to disassociate semantics from syntax. Our work further focuses on finding the corresponding answers from the target language based on better knowledge transfer and textual alignments from multiple granularities.

3 X-STA: Proposed Approach

In this section, we present the detailed techniques of X-STA for cross-lingual MRC.

3.1 Task Definition and Basic Notations

Given a context \( C \) and a question \( Q \), the MRC task is to extract a sub-sequence from context \( C \) as the right answer to question \( Q \). Denote the input sequence as \( X = \{Q, C\} \in \mathbb{R}^N \), where \( N \) is the sequence length. We use \( p_{\text{start}} \in \mathbb{R}^N \) and \( p_{\text{end}} \in \mathbb{R}^N \) to denote the answer start and end position probability distributions. For the sake of simplicity, we concatenate the two together to \( p \in \mathbb{R}^{N \times 2} \). Similarly, \( y \in \mathbb{R}^{N \times 2} \) represents the one-hot golden label sequence. For cross-lingual scenarios, only annotated training data from the source language \( D_{\text{Train}}^S = \{X^S_{\text{Train}}, y^S_{\text{Train}}\} \) and raw test data from the target language \( D_{\text{Test}}^T = \{X^T_{\text{Test}}, y^T_{\text{Test}}\} \) are available. \( S \) and \( T \) denote the source and target language.

Machine translation can be used to obtain training data for the target language \( D_{\text{Train}}^T = \{X^T_{\text{Train}}, y^T_{\text{Train}}\} \) and test data for the source language \( D_{\text{Test}}^S = \{X^S_{\text{Test}}, y^S_{\text{Test}}\} \) (Hu et al., 2020). In addition, we use \( h^l \) to denote the hidden states of a sequence in layer \( l \in L \), where \( L \) is the total number of transformer layers. Thus, to predict the start position and end position of the correct answer span in \( X \), the probability distributions \( p \) is induced over the entire sequence by feeding \( h^L \) into a linear classification layer and a softmax function: \( p = \text{softmax}(W \cdot h^L + b) \), \( W \) and \( b \) are the weights and bias of the linear classifier.

![Figure 2: Model architecture. The cross-attention block (GDKS) is implemented only in a certain layer. In other layers, vanilla transformer layers are applied.](image)

3.2 Gradient-Disentangled Knowledge Sharing

Although machine translation from high-resource languages to low-resource ones can be used for training multi-lingual models, the drawbacks are evident. i) Machine translation quality varies across languages. ii) The original semantics can be easily lost during translation. iii) Task labels are relatively expensive to obtain, especially for token-level cross-lingual tasks. Thus, as shown in 2, we leverage parallel language pairs as the input and fuse cross-lingual representations.

As in Yang et al. (2022), cross-attention can be leveraged for feature fusion. However, a performance loss can be observed in the source language, as shown in Figure 3. A reasonable conjecture is that helping the target language to extract target-related information from the hidden states of the source language leads to a degeneration of source language representations. To alleviate this problem, we propose Gradient-Disentangled Knowledge Sharing (GDKS), which is an improved version of the cross-attention block. Specifically, we
block the gradients from the target language hidden states back to the source language hidden states $h^S_t$. As a compensation, we add a trainable correction term:

$$\tilde{h}^S_t = \begin{cases} h^S_t & \text{if } S = T \\ s_{\text{g}}(h^S_t) + f(s_{\text{g}}(h^S_t), s_{\text{g}}(h^T_t)) & \text{otherwise} \end{cases}$$

Here, $s_{\text{g}}(\cdot)$ is used to stop back-propagating gradients, preventing interfering with source language representations. $f(\cdot)$ refers to a trainable linear transformation with dropout. Then use the target hidden states as the query and the converted source hidden states $\tilde{h}^S_t$ as key and value to perform cross-attention, defined as follows:

$$h^T_{T|S} = \text{MHA}(h^T_T, \tilde{h}^S_t, \tilde{h}^T_t)$$

where MHA is multi-head attention (Vaswani et al., 2017). Then, the target hidden states are fused with the source-aware target hidden states by the weight $\lambda$, computed as follows:

$$h^{T+1}_T = (1 - \lambda) \cdot h^T_{T|S} + \lambda \cdot \text{MHA}(h^T_T, h^T_T, h^T_T)$$

where $\lambda = w \cdot \lambda_0 + b$, with $w$ and $b$ to be trainable parameters. It is worth noting that GDKS is implemented in a certain transformer layer only.

### 3.3 Attentive Teacher-Guided Calibration

As GDKS focuses on transferring knowledge from hidden states of the teacher model (trained from the source language), we also calibrate the model output distributions with teacher guidance.

**Normalization.** The premise of obtaining good guidance is that the representations of different languages should be normalized first. Following Pires et al. (2019), we hypothesize that the representation of a multi-lingual model is composed of language-specific and language-agnostic representations. We estimate language-specific features as the mean of the language representations and remove language-specific features by subtracting the mean to retain only the generic semantic features. The intuition behind this is that a certain language may have a large number of phenomena such as function words (Libovicky et al., 2020). Therefore, the average representation of that language is prominent. Inspired by Batch Normalization (Ioffe and Szegedy, 2015), we transform the generic semantic representation to the standard normal distribution space:

$$h = \frac{h^L - \mu_\beta}{\sqrt{\beta^2 + \epsilon}}$$

where $\mu_\beta$ and $\sigma_\beta$ are mean and variance of token-level representations in batch $\beta$. $\epsilon$ is a constant for numerical stability. To facilitate its use in inference, it is set to be linguistically independent.

**Calibration.** After normalization, we use the hidden states of the target language as query, and the hidden states and the output distribution of the source language as key and value, respectively. We also leverage MHA and average the results of the transformation of multiple heads. Hence, the transferred output distribution $p_{T|S} \in \mathbb{R}^{N \times 2}$ is:

$$p_{T|S} = \text{MHA}(h^T_T, h^S_t, s_{\text{g}}(p_S))$$

where $h^T_T$ and $h^S_t$ are the normalized hidden states of source and target languages, respectively.

During the model training phase, we incorporate a teacher-guided loss $L_{tg}$ for the computation of $p_{T|S}$. Thus, tokens with the same semantics but in different languages can still be brought closer together by annotated data, even if their representations differ significantly. Specifically, we have the sample-wise loss $L_{tg}$ defined as follows:

$$L_{tg} = -\sum_i \sum_j y_i^j \log p_i^j.$$
3.4 Multi-Granularity Semantic Alignment

We further enhance the knowledge transfer of our model, based on our proposed Multi-Granularity Semantic Alignment (MGSA) technique.

Sentence-Level Alignment. A vanilla approach to learn alignments is from the sentence level. Here, we employ Contrastive Learning (CL, Hadlens et al., 2006; Chen et al., 2020) to strengthen the alignment across languages:

$$L_{\text{align}} = -\log \frac{e^{\text{sim}(r,r^+)/\tau}}{e^{\text{sim}(r,r^+)/\tau} + \sum_i e^{\text{sim}(r,r_i^+)/\tau}}$$

where $r$ is the mean pooled sentence representation, $r^+$ and $r^-$ represent a positive sample from the parallel translated data and a negative example in the mini-batch, respectively. sim$(r_1, r_2)$ is the cosine similarity, i.e., sim$(r_1, r_2) = \frac{r_1^T r_2}{\|r_1\| \|r_2\|}$. $\tau$ is the temperature hyper-parameter, which we set to 0.05 in default.

Token-Level Alignment. In Fomicheva et al. (2020); Yang et al. (2022), the entropy of the cross-attention distribution (ECA) is used to measure the quality of machine translation. A smaller entropy of the attention distribution, i.e., more focused attention, can indicate a relatively higher translation quality (Rikters and Fishel, 2017). Similarly, ECA can also be used to represent the cross-lingual alignment quality, which we use as a penalty term for training the cross-lingual model to avoid distraction in GDKS. The token-level alignment loss $L_{\text{align}_T}$ can be defined as:

$$L_{\text{align}_T} = -\frac{1}{T} \sum_i \sum_j a_{ij} \log a_{ij}$$

where $a_{ij} = \text{softmax}(\frac{h_i^r h_j^s}{\sqrt{n}})$ represents attention weights, $n$ is the hidden size, $T$ is the number of target tokens and $J$ is the number of source tokens. Next, the total alignment loss is summed by the two parts, with $\varsigma$ and $\eta$ to be the coefficients:

$$L_{\text{align}} = \varsigma L_{\text{align}_S} + \eta L_{\text{align}_T}.$$  

3.5 Final Training Objective

In brief, the final training objective of X-STA is:

$$L = L_{\text{MRC}} + \gamma L_{tg} + L_{\text{align}}$$

where $\gamma$ is a factor for the teacher-guided loss $L_{tg}$. $L_{\text{MRC}}$ refers to the cross-entropy loss of the MRC task. Following Yang et al. (2022), we split the MRC loss $L_{\text{MRC}}$ into the MRC loss of the source language and the target language with a balancing factor $\alpha$:

$$L_{\text{MRC}} = \alpha L_{\text{MRC}}^s + (1 - \alpha) L_{\text{MRC}}^t.$$  

4 Experiments

4.1 Datasets

We evaluate X-STA on three multi-lingual MRC datasets, namely MLQA (Lewis et al., 2020), XQuAD (Artetxe et al., 2020) and TyDiQA (Clark et al., 2020). MLQA is a benchmark dataset consisting of over 5K extractive MRC instances in 7 languages: English (en), Arabic (ar), German (de), Spanish (es), Hindi (hi), Russian (ru), Thai (th), Turkish (tr), Vietnamese (vi), and Chinese (zh). XQuAD consists of a subset of 240 paragraphs and 1190 question-answer pairs from the SQuAD v1.1 (Rajpurkar et al., 2016) development set together with their professional translations into ten languages: English (en), Arabic (ar), German (de), Greek (el), Spanish (es), Hindi (hi), Russian (ru), Thai (th), Turkish (tr), Vietnamese (vi), and Chinese (zh). TyDiQA covers 9 typologically diverse languages: English (en), Arabic (ar), Bengali (bn), Finnish (fi), Indonesian (id), Korean (ko), Russian (ru), Swahili (sw), Telugu (te). Follow XTREME (Hu et al., 2020), we use the gold passage version of TyDiQA.

For the translated data, we employ the translate-train and translate-test data from XTREME 2. We use two evaluation metrics, namely exact match (EM) and macro-average F1 score (F1), following Rajpurkar et al. (2016); Hu et al. (2020).

4.2 Experimental Settings

We conduct extensive experiments based on two multi-lingual pre-trained backbones: mBERT (Devlin et al., 2019) and XLM-R-base (Conneau et al., 2020). The batch size is set to 32. The learning rate is set to 3e-5, and decreases linearly with warmup. Following Yang et al. (2022), $\alpha$ is set to 0.2 and we implement GDKS in the 8th layer. We set $\lambda_0$ to 0.3 and $\epsilon$ to 1e-8. We perform grid search $\varsigma$, $\eta$ and $\gamma$ from [0.01, 0.05, 0.1, 0.5] on the validation set of MLQA, and finally set them to 0.05, 0.05 and 0.1, respectively. We save the model with the best averaged performance of all languages on the validation set for testing. Since there are no validation sets in XQuAD and TyDiQA. Following

2https://github.com/google-research/xtreme
Table 1: Overall evaluation (F1/EM) over the MLQA dataset. * denotes the results of our re-implementation.

<table>
<thead>
<tr>
<th>Methods</th>
<th>en</th>
<th>ar</th>
<th>de</th>
<th>es</th>
<th>hi</th>
<th>vi</th>
<th>zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on mBERT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-shot</td>
<td>80.2/67.0</td>
<td>52.3/34.6</td>
<td>59.0/43.8</td>
<td>67.4/49.2</td>
<td>50.2/35.3</td>
<td>61.2/40.7</td>
<td>59.6/38.6</td>
<td>61.4/44.2</td>
</tr>
<tr>
<td>Trans-train</td>
<td>80.7/67.7</td>
<td>58.9/39.0</td>
<td>66.0/51.6</td>
<td>71.3/53.7</td>
<td>62.4/45.0</td>
<td>67.9/47.6</td>
<td>66.0/43.9</td>
<td>67.6/49.8</td>
</tr>
<tr>
<td>LAKM</td>
<td>80.1/66.9</td>
<td>-</td>
<td>64.4/49.9</td>
<td>69.5/51.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>69.0/50.9</td>
</tr>
<tr>
<td>Ours</td>
<td>80.5/67.6</td>
<td>64.1/43.7</td>
<td>69.2/54.6</td>
<td>74.2/56.5</td>
<td>67.6/49.7</td>
<td>73.5/52.8</td>
<td>69.2/47.7</td>
<td>71.2/53.2</td>
</tr>
</tbody>
</table>

Based on XLM-R

<table>
<thead>
<tr>
<th>Methods</th>
<th>en</th>
<th>ar</th>
<th>de</th>
<th>es</th>
<th>hi</th>
<th>vi</th>
<th>zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-shot</td>
<td>79.2/66.2</td>
<td>56.2/37.2</td>
<td>61.7/46.9</td>
<td>67.4/50.0</td>
<td>61.5/44.2</td>
<td>65.6/45.2</td>
<td>62.5/39.0</td>
<td>64.9/47.0</td>
</tr>
<tr>
<td>Trans-train</td>
<td>80.9/67.9</td>
<td>59.8/40.4</td>
<td>65.2/50.8</td>
<td>70.3/52.9</td>
<td>65.1/47.9</td>
<td>69.3/49.1</td>
<td>63.4/41.3</td>
<td>67.7/50.1</td>
</tr>
<tr>
<td>CalibreNet</td>
<td>79.7/66.6</td>
<td>56.1/37.8</td>
<td>61.7/47.6</td>
<td>68.0/50.8</td>
<td>60.0/43.8</td>
<td>66.9/46.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AA-CL</td>
<td>80.1/66.8</td>
<td>58.5/41.3</td>
<td>64.6/49.8</td>
<td>69.0/51.2</td>
<td>62.8/46.5</td>
<td>67.9/47.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>78.9/65.8</td>
<td>62.5/43.1</td>
<td>65.7/51.5</td>
<td>71.8/54.5</td>
<td>66.8/49.6</td>
<td>71.4/50.9</td>
<td>65.3/43.4</td>
<td>68.9/51.3</td>
</tr>
<tr>
<td>Ours</td>
<td>81.6/68.7</td>
<td>63.1/43.2</td>
<td>67.5/52.9</td>
<td>72.7/55.1</td>
<td>68.6/50.9</td>
<td>72.7/52.0</td>
<td>68.6/43.5</td>
<td>70.4/52.3</td>
</tr>
</tbody>
</table>

4.3 Baselines

We systematically compare our method with the following strong baselines:

- **Zero-shot** models are trained on labeled data in the source language only, and directly evaluated on target languages.

- **Trans-train** (Hu et al., 2020) translates training data in English into target languages. The model is trained on the combination of these original and translated training sets.

- **LAKM** (Yuan et al., 2020) leverages a language-agnostic knowledge masking task by knowledge phrases based on mBERT.

- **CalibreNet** (Liang et al., 2021) employs a unsupervised phrase boundary recovery pre-training task to enhance the multi-lingual boundary detection capability of XLM-R_{base}.

- **AA-CL** (Chen et al., 2022) is a two-stage step-by-step algorithm for finding the best answer for cross-lingual MRC over XLM-R_{base}.

- **X-MIXUP** (Yang et al., 2022) is a cross-lingual manifold mixup method that learns compromised representations for target languages, which produces the state-of-the-art results for cross-lingual MRC.

4.4 General Experimental Results

For Zero-shot and Trans-train, we report the results of mBERT from Hu et al. (2020) and reproduce the results of XLM-R_{base}. For LAKM, CalibreNet and AA-CL (which have been evaluated over part of our settings), we report the results from their original papers. As for X-MIXUP (the state-of-the-art method), in order to conduct a rigorous comparison, we report both the results from the original paper and our re-implementation. Among these methods, only Zero-shot and CalibreNet are under zero-shot setting, for the rest of the methods translate data are available.

For X-MIXUP significantly reduces the performance gap between the source and target languages, but also compromises performance on the source language; whereas our approach can achieve comparable performance to translate-train on English without negatively affecting the representation of the source language.

Yang et al. (2022), for the former, we use MLQA’s validation set and for the latter, we use the English data as the validation set. All the experiments are implemented in PyTorch and run on a single server with NVIDIA Tesla V100 (32GB) GPUs.
### Table 2: Overall evaluation (F1/EM) over the TyDiQA dataset. * denotes the results of our re-implementation.

<table>
<thead>
<tr>
<th>Methods</th>
<th>en</th>
<th>ar</th>
<th>bn</th>
<th>fi</th>
<th>id</th>
<th>ko</th>
<th>ru</th>
<th>sw</th>
<th>te</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on mBERT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-shot</td>
<td>75.3/63.6</td>
<td>62.2/42.8</td>
<td>49.3/32.7</td>
<td>59.7/45.3</td>
<td>64.8/45.8</td>
<td>58.8/50.0</td>
<td>60.0/38.8</td>
<td>57.5/37.9</td>
<td>49.6/38.4</td>
<td>59.7/43.9</td>
</tr>
<tr>
<td>Trans-train</td>
<td>73.2/62.5</td>
<td>71.8/54.2</td>
<td>49.7/36.3</td>
<td>68.1/53.6</td>
<td>72.3/55.2</td>
<td>58.6/47.8</td>
<td>64.3/45.3</td>
<td>66.8/48.9</td>
<td>53.3/40.2</td>
<td>64.2/49.3</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>60.8/46.5</td>
</tr>
<tr>
<td>X-MIXUP*</td>
<td>72.5/60.7</td>
<td>70.0/52.8</td>
<td>55.1/41.6</td>
<td>65.8/50.0</td>
<td>74.1/57.7</td>
<td>62.6/52.2</td>
<td>63.0/43.3</td>
<td>67.5/49.1</td>
<td>51.2/37.5</td>
<td>64.6/49.4</td>
</tr>
<tr>
<td>Ours</td>
<td>73.9/63.4</td>
<td>72.4/54.5</td>
<td>60.9/47.8</td>
<td>69.4/55.9</td>
<td>76.2/60.9</td>
<td>64.0/52.2</td>
<td>65.2/46.0</td>
<td>71.2/54.1</td>
<td>51.2/41.7</td>
<td>67.2/52.9</td>
</tr>
<tr>
<td>Based on XLM-R_{base}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-shot*</td>
<td>66.0/53.4</td>
<td>61.1/41.6</td>
<td>37.8/23.0</td>
<td>61.4/45.7</td>
<td>72.6/55.0</td>
<td>48.1/33.0</td>
<td>59.5/35.0</td>
<td>54.7/35.9</td>
<td>37.5/25.4</td>
<td>55.4/38.7</td>
</tr>
<tr>
<td>Trans-train*</td>
<td>70.9/59.2</td>
<td>67.7/49.6</td>
<td>46.3/31.0</td>
<td>65.1/51.3</td>
<td>74.2/57.5</td>
<td>54.3/43.1</td>
<td>63.9/46.0</td>
<td>63.2/47.1</td>
<td>63.3/46.9</td>
<td>63.2/48.0</td>
</tr>
<tr>
<td>X-MIXUP*</td>
<td>68.0/54.8</td>
<td>67.7/48.8</td>
<td>50.6/33.6</td>
<td>66.5/52.6</td>
<td>72.0/55.0</td>
<td>52.7/40.6</td>
<td>64.0/45.0</td>
<td>64.0/47.5</td>
<td>60.2/43.3</td>
<td>62.9/46.8</td>
</tr>
<tr>
<td>Ours</td>
<td>71.3/59.3</td>
<td>68.6/50.8</td>
<td>56.7/40.7</td>
<td>67.6/54.1</td>
<td>77.7/62.5</td>
<td>55.7/44.6</td>
<td>64.2/46.1</td>
<td>64.6/48.5</td>
<td>70.2/52.9</td>
<td>66.3/51.0</td>
</tr>
</tbody>
</table>

4.5 Ablation Study

We conduct an ablation study by removing each key component individually to evaluate the effectiveness of our method. As shown in Table 4, there is a performance gap when removing any of the components. Although the removal of GDKS has little effect on the overall performance, it significantly affects the representation of the source language, resulting in an obvious performance drop in the source language (i.e., English).

Removing the Attentive Teacher-Guided Calibration (ATGC) component degrades the model performance the most, and the results demonstrate that mapping the output of the source language to the target language space is efficient and feasible. In addition, there is still some performance loss compared to removing ATGC only at inference time, which suggests that the improvement from ATGC does not only come from weighted outputs of source and target languages. Using the answer span as additional knowledge can enhance the cross-lingual alignment through the teacher-guided loss \( L_{tg} \). Figure 5 shows the visual distribution of sentence representations before and after normalization. In the original space, the distributions of the same language (same color) tend to cluster together, while after normalization, these representations are sparsely dispersed, which also shows that normalization can indeed decompose some of the language-specific representations.

Finally, we analyze the effectiveness of MGSA. As seen, token-level alignment contributes more than sentence-level alignment. A reasonable speculation is that the MRC task is more concerned with token-level representations. As shown in Figure 4, without token-level alignment, attention is
Based on mBERT

Zero-shot 83.5/72.2 61.5/45.1 70.6/54.0 62.6/44.9 75.5/56.9 59.2/46.0
Trans-train 86.0/74.5 71.0/54.1 78.8/63.9 74.2/56.1 82.4/66.2 71.3/56.2
X-MIXUP - - - - - -
X-MIXUP* 83.4/71.9 78.0/60.9 80.6/65.2 79.0/60.7 81.7/63.7 77.4/61.8
Ours 85.6/74.4 80.1/62.9 82.3/66.6 81.3/64.4 83.5/65.0 79.0/64.2

Based on XLM-R base

Zero-shot 71.3/53.3 42.7/33.5 55.4/40.1 69.5/49.6 58.0/48.3 64.5/49.4
Trans-train 78.1/63.0 38.1/34.5 70.6/55.7 78.5/58.8 67.7/58.7 72.4/58.3
AA-CL - - - - - -
X-MIXUP* 80.1/63.9 61.9/56.6 75.0/58.2 80.4/61.1 71.2/61.8 77.1/62.3
Ours 81.8/66.2 65.2/59.4 76.8/62.2 82.0/63.8 70.2/60.3 78.9/64.5

Based on XLM-Rbase

Zero-shot* 74.4/58.8 64.6/53.4 67.7/51.1 73.7/53.6 61.4/52.4 71.3/56.1
Trans-train* 77.0/61.3 59.9/55.0 72.2/56.8 77.3/59.1 74.9/72.3 75.1/61.5
AA-CL - - - - - -
X-MIXUP* 77.8/62.1 72.7/61.1 72.2/56.5 78.2/59.2 77.4/73.9 76.6/62.6
Ours 80.8/66.0 70.2/64.5 76.4/61.2 81.0/63.3 77.7/74.6 79.4/65.8

Table 3: Overall evaluation (F1/EM) over the XQuAD dataset. * denotes the results of our re-implementation.

Ablation MLQA XQuAD
Ours 71.2 / 53.2 78.9 / 64.5
w/o. GDKS 71.1 / 53.0† 78.0 / 63.5‡
w/o. ATGC 70.8 / 52.6 77.4 / 63.1
w/o. ATGC inference 70.9 / 53.0 78.6 / 64.2
w/o. Rep-Norm 71.0 / 52.9 78.0 / 63.5
w/o. alignment 71.2 / 53.0 78.5 / 64.0
w/o. alignment 70.8 / 52.6 77.5 / 63.4

Table 4: Ablation study of our method on MLQA and XQuAD. w/o. GDKS refers to vanilla cross-attention is used, Rep-Norm is Representation Normalization. alignment, and alignment, refer to sentence-level alignment and token-level alignment. † and ‡ have a performance drop of 1.1/1.0 and 1.3/1.0 on English.

4.6 Case Study

To further demonstrate the output results of our approach, we show the answer generation process of a Hindi example and the corresponding English example from the XQuAD dataset, and compare it with a powerful ultra-large language model (i.e., ChatGPT). Figure 6 shows that for English, both our approach and ChatGPT answer the question well. However, in a low-resource language setting such as Hindi, there are some capability limitations of mBERT and ChatGPT. Without ATGC, our method fails to find the correct answer. When mapping the source language output to the target language output space, it successfully calibrates the output and generates the correct answer after averaging the two outputs. ChatGPT, on the other hand, produces plausible but incorrect answers, showing a sign of producing hallucinations (also reported in Bang et al. (2023)). More cases in low-resource languages can be found in Appendix B.

5 Conclusion

In this paper, we propose X-STA, which addresses the challenges of cross-lingual MRC in effectively utilizing translation data and the linguistic and cultural differences. Our work follows three principles: sharing, teaching and aligning. Experimental results on three datasets show that our approach obtains the state-of-the-art performance compared to strong baselines. We further analyze the effec-
tiveness of each component. In the future, we will extend our work to other cross-lingual NLP tasks for low-resource languages.

Limitations

Our approach requires a translation system as an aid and incurs additional inference costs during the inference process (the sequences translated back to the source language also need to go through model). For other cross-lingual token-level tasks (e.g., POS, NER), it is difficult to obtain the labels of translate-train data directly. Previous approaches usually use trained models to generate pseudo-labels. These low-quality labels pose significant challenges to our approach. Extending our approach to these tasks is left to our subsequent work.

Acknowledgements

This work is partially supported by Alibaba Cloud Group, through Research Talent Program with South China University of Technology.

References


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


Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r:


A Parameter Analysis

To further evaluate the effectiveness of ATGC, we conduct a series of experiments on MLQA and XQuAD for the hyper-parameter $\gamma = \{0, 0.01, 0.05, 0.1, 0.5\}$. Figure 7 shows that the performance trends are consistent over both datasets, with $\gamma$ achieving optimal performance on 0.1. We conjecture that when $\gamma$ is too large, it can interfere with the original representation of the model. In addition, we conduct the implementation of GDKS in different layers and find that the optimal number of layers to implement is 8, which is consistent with the results of X-MIXUP (Yang et al., 2022).

B Cases

As shown in Figure 9, 10 and 11, ChatGPT sometimes fail to generate the answers accurately. In contrast, our method is able to extract the correct answers with the help of the English datasets.

Figure 7: The parameter analysis of $\gamma$ on MLQA (left vertical axis) and XQuAD (right vertical axis).

Figure 8: Performance comparison on implementing GDKS in different layers over the MLQA dataset.
Passage: As northwest Europe slowly began to warm up from 22,000 years ago onward, frozen subsoil and expanded alpine glaciers began to thaw and fall-winter snow covers melted in spring. Much of the discharge was routed to the Rhine and its downstream extension. Rapid warming and changes of vegetation, to open forest, began about 13,000 BP. By 9000 BP, Europe was fully forested. With globally shrinking ice-cover, ocean water levels rose and the English Channel and North Sea re-inundated. Meltwater, adding to the ocean and land subsidence, drowned the former coasts of Europe transgressively.

Question: when was Europe fully forested and recovered from the last Ice Age?

Ours: 9000 BP
ChatGPT: Europe was fully forested and recovered from the last Ice Age by 9000 BP.

Figure 9: A Thai (th) example from the XQuAD dataset.

Passage: Through combining the definition of electric current as the time rate of change of electric charge, a rule of vector multiplication called Lorentz's law describes the force on a charge moving in a magnetic field. The connection between electricity and magnetism allows for the description of a unified electromagnetic force that acts on a charge. This force can be written as a sum of the electrostatic force (due to the electric field) and the magnetic force (due to the magnetic field). Fully stated, this is the law:

Question: what magnetic and electric force acts on a charge?

Ours: unified electromagnetic force
ChatGPT: The magnetic force and electrostatic force act on a charge.

Figure 10: A Vietnamese (vi) example from the XQuAD dataset.

Passage: Reserved matters are subjects that are outside the legislative competence of the Scotland Parliament. The Scottish Parliament is unable to legislate on such issues that are reserved to, and dealt with at, Westminster (and where Ministerial functions usually lie with UK Government ministers). These include abortion, broadcasting policy, civil service, common markets for UK goods and services, constitution, electricity, coal, oil, gas, nuclear energy, defence and national security, drug policy, employment, foreign policy and relations with Europe, most aspects of transport safety and regulation, National lottery, protection of borders, social security and stability of UK’s fiscal, economic and monetary system.

Question: Where are issues like abortion and drug policy legislated on?

Ours: Westminster
ChatGPT: Europe was fully forested and recovered from the last Ice Age by 9000 BP.

Figure 11: A Chinese (zh) example from the XQuAD dataset.

Passage: 保留问题是苏格兰议会立法能力以外的议题。苏格兰议会无法对保留问题进行立法，应交由威斯敏斯特处理（部委职能通常由英国政府的国务大臣完成。此类问题包括堕胎、广播政策、国民政府、政府机构、英国商品和服务的共同市场、宪法、电力、煤炭、石油、核电、核能、国防和国家安全、毒品政策、就业、对外政策和对欧关系、交通安全和管理的大多数方面、英国彩票、边境安全、社会保障以及英国的财政、经济和货币政策的稳定性。

Question: 诸如堕胎、毒品政策之类的问题在哪里立法?

Ours: 威斯敏斯特
ChatGPT: 这些问题的立法权不在苏格兰议会，而是应交由威斯敏斯特处理，通常由英国政府的国务大臣完成。