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

Tingfeng Cao[1,2,3]*, **Chengyu Wang**[2]†, **Chuanqi Tan**[2], **Jun Huang**[2], **Jinhui Zhu**[1,3]†
[1]School of Software Engineering, South China University of Technology, China
[2]Alibaba Group, China
[3]Key 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]

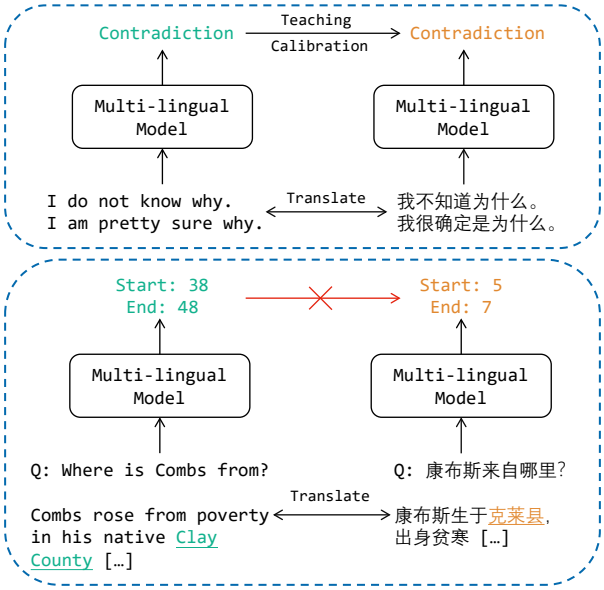

Figure 1: Machine translation as an aid for cross-lingual transfer. Above is a natural language inference (NLI) task. The probability distribution of the source language can be fitted by KL Divergence for teaching low-resource languages; during inference, the target language can be translated into the source language with its output used for calibration. Below is an MRC task, where the knowledge is difficult to transfer directly.

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

*Work done during an internship at Alibaba.
†C. Wang and J. Zhu are co-corresponding authors.
[1]Source codes will be publicly available in the EasyNLP framework (Wang et al., 2022). URL: https://github.com/alibaba/EasyNLP.

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 455–467
December 6-10, 2023

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 mainfold 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 the 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 $\mathbf{X} = \{Q, C\} \in \mathbb{R}^N$, where N is the sequence length. We use $\mathbf{p}_{\text{start}} \in \mathbb{R}^N$ and $\mathbf{p}_{\text{end}} \in \mathbb{R}^N$ to denote the answer start and end position probability distributions. For the sake of simplicity, we concatenate the two together to $\mathbf{p} \in \mathbb{R}^{N\times 2}$. Similarly, $\mathbf{y} \in \mathbb{R}^{N\times 2}$ represents the one-hot golden label sequence. For cross-lingual scenarios, only annotated training data from the source language $D_S^{\text{Train}} = \{\mathbf{X}_S^{\text{Train}}, \mathbf{y}_S^{\text{Train}}\}$ and raw test data from the target language $D_T^{\text{Test}} = \{\mathbf{X}_T^{\text{Test}}, \mathbf{y}_T^{\text{Test}}\}$ are available. S and T denote the source and target language. Machine translation can be used to obtain training data for the target language $D_T^{\text{Train}} = \{\mathbf{X}_T^{\text{Train}}, \mathbf{y}_T^{\text{Train}}\}$ and test data for the source language $D_S^{\text{Test}} = \{\mathbf{X}_S^{\text{Test}}, \mathbf{y}_S^{\text{Test}}\}$ (Hu et al., 2020). In addition, we use $\mathbf{h}^l$ 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 $\mathbf{X}$, the probability distributions $\mathbf{p}$ is induced over the entire sequence by feeding $\mathbf{h}^L$ into a linear classification layer and a softmax function: $\mathbf{p} = \text{softmax}(\mathbf{W}\mathbf{h}^L + \mathbf{b})$, $\mathbf{W}$ and $\mathbf{b}$ are the weights and bias of the linear classifier.

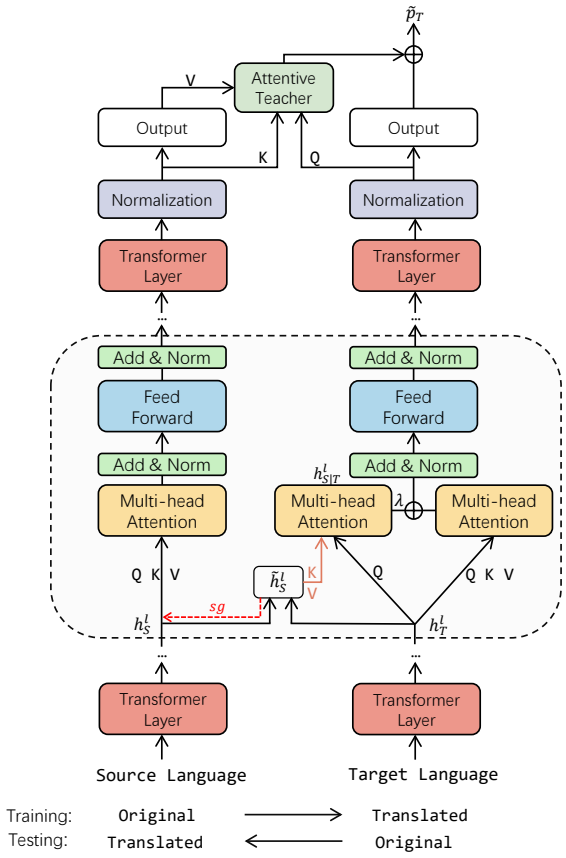

Figure 2: Model architecture. The cross-attention block (GDKS) is implemented only in a certain layer. In other layers, vanilla transformer layers are applied.

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

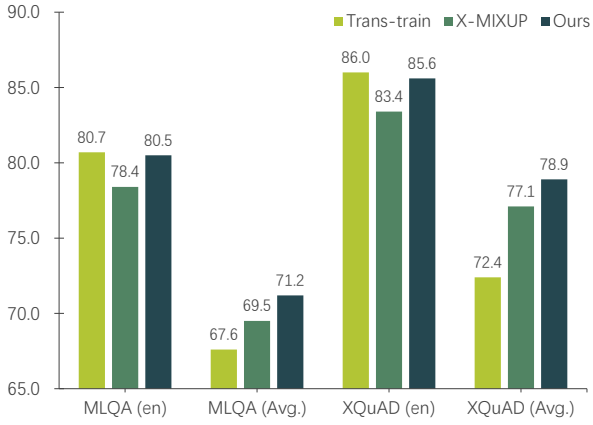

Figure 3: The performance of previous methods and our method on cross-lingual MRC. X-MIXUP (a cross-attention based approach) improves the performance on target languages, but with a performance drop on the source language (en). Our approach addresses the issue by GDKS.

block the gradients from the target language output back to the source language hidden states $\mathbf{h}_S^l$. As a compensation, we add a trainable correction term:

$$\tilde{\mathbf{h}_S^l} = \begin{cases} \mathbf{h}_S^l & S = T \\ sg(\mathbf{h}_S^l) + f(sg(\mathbf{h}_S^l), sg(\mathbf{h}_T^l)) & \text{otherwise} \end{cases}$$

Here, $sg(\cdot)$ 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 $\tilde{h_S^l}$ as key and value to perform cross-attention, defined as follows:

$$\mathbf{h}_{T|S}^l = \text{MHA}(\mathbf{h}_T^l, \tilde{\mathbf{h}_S^l}, \tilde{\mathbf{h}_S^l})$$

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 λ, computed as follows:

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

where $\lambda = w * \lambda_0 + b$, 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 (Libovický 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:

$$\tilde{\mathbf{h}} = \frac{\mathbf{h}^L - \boldsymbol{\mu}_\beta}{\sqrt{\boldsymbol{\sigma}_\beta^2 + \epsilon}}$$

where $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\sigma}_\beta$ are mean and variance of token-level representations in batch β. ϵ 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 $\mathbf{p}_{T|S} \in \mathbb{R}^{N \times 2}$ is:

$$\mathbf{p}_{T|S} = \text{MHA}(\tilde{\mathbf{h}_T}, \tilde{\mathbf{h}_S}, sg(\mathbf{p}_S))$$

where $\tilde{\mathbf{h}_T}$ and $\tilde{\mathbf{h}_S}$ are the normalized hidden states of source and target languages, respectively.

During the model training phase, we incorporate a teacher-guided loss $\mathcal{L}_{tg}$ for the computation of $\mathbf{p}_{T|S}$. 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 $\mathcal{L}_{tg}$ defined as follows:

$$\mathcal{L}_{tg} = -\sum_{i}^{N}\sum_{j}^{2} \mathbf{y}^{ij} log \mathbf{p}_{T|S}^{ij}.$$

For model inference, we leverage $\mathbf{p}_{T|S}$ to calibrate the output for the target language by averaging the results from two output distributions, i.e., $\tilde{\mathbf{p}_T} = \frac{\mathbf{p}_{T|S} + \mathbf{p}_T}{2}$.

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, Hadsell et al., 2006; Chen et al., 2020) to strengthen the alignment across languages:

$$\mathcal{L}_{align_S} = -log\frac{e^{\text{sim}(\mathbf{r},\mathbf{r}^+)/\tau}}{e^{\text{sim}(\mathbf{r},\mathbf{r}^+)/\tau} + \sum_i e^{\text{sim}(\mathbf{r},\mathbf{r}_i^-)/\tau}}$$

where $\mathbf{r}$ is the mean pooled sentence representation. $\mathbf{r}^+$ and $\mathbf{r}^-$ represent a positive sample from the parallel translated data and a negative example in the mini-batch, respectively. $\text{sim}(\mathbf{r}_1, \mathbf{r}_2)$ is the cosine similarity, i.e., $\text{sim}(\mathbf{r}_1, \mathbf{r}_2) = \frac{\mathbf{r}_1^\top \mathbf{r}_2}{\|\mathbf{r}_1\| \cdot \|\mathbf{r}_2\|}$. τ 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 $\mathcal{L}_{align_T}$ can be defined as:

$$\mathcal{L}_{align_T} = -\frac{1}{I}\sum_{i}^{I}\sum_{j}^{J} a^{ij} log a^{ij}$$

where $a^{ij} = \text{softmax}(\frac{\mathbf{h}_{T_i}\mathbf{h}_{S_j}^\top}{\sqrt{n}})$ represents attention weights, n is the hidden size, I 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 ς and η to be the coefficients:

$$\mathcal{L}_{align} = \varsigma\, \mathcal{L}_{align_S} + \eta\, \mathcal{L}_{align_T}.$$

3.5 Final Training Objective

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

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

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

$$\mathcal{L}_{\text{MRC}} = \alpha\mathcal{L}^S_{\text{MRC}} + (1-\alpha)\mathcal{L}^T_{\text{MRC}}.$$

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), 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), α is set to 0.2 and we implement GDKS in the 8th layer. We set λ_0 to 0.3 and ϵ to 1e-8. We perform grid search ς, η and γ 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

[2] https://github.com/google-research/xtreme

Methods	en	ar	de	es	hi	vi	zh	Avg.
Based on mBERT								
Zero-shot	80.2/67.0	52.3/34.6	59.0/43.8	67.4/49.2	50.2/35.3	61.2/40.7	59.6/38.6	61.4/44.2
Trans-train	**80.7/67.7**	58.9/39.0	66.0/51.6	71.3/53.7	62.4/45.0	67.9/47.6	66.0/43.9	67.6/49.8
LAKM	80.1/66.9	-	64.4/49.9	69.5/51.5	-	-	-	-
X-MIXUP	-	-	-	-	-	-	-	69.0/50.9
X-MIXUP*	78.4/64.9	63.3/43.5	67.5/53.6	72.3/55.0	65.8/47.5	72.2/51.7	66.6/45.6	69.5/51.7
Ours	80.5/67.6	**64.1/43.7**	**69.2/54.6**	**74.2/56.5**	**67.6/49.7**	**73.5/52.8**	**69.2/47.7**	**71.2/53.2**
Based on XLM-R_{base}								
Zero-shot*	79.2/66.2	56.2/37.2	61.7/46.9	67.4/50.0	61.5/44.2	65.6/45.2	62.5/39.0	64.9/47.0
Trans-train*	80.9/67.9	59.8/40.4	65.2/50.8	70.3/52.9	65.1/47.9	69.3/49.1	63.4/41.3	67.7/50.1
CalibreNet	79.7/66.6	56.1/37.8	61.7/47.6	68.0/50.8	60.0/43.8	66.9/46.6	-	-
AA-CL	80.1/66.8	58.5/41.3	64.6/49.8	69.0/51.2	62.8/46.5	67.9/47.2	-	-
X-MIXUP*	78.9/65.8	62.5/43.1	65.7/51.5	71.8/54.5	66.8/49.6	71.4/50.9	65.3/43.4	68.9/51.3
Ours	**81.6/68.7**	**63.1/43.2**	**67.5/52.9**	**72.7/55.1**	**68.6/50.9**	**72.7/52.0**	**66.3/43.5**	**70.4/52.3**

Table 1: Overall evaluation (F1/EM) over the MLQA dataset. * denotes the results of our re-implementation.

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.

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.

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.

4.4 General Experimental Results

As in Table 1, based on mBERT, we achieved an average of 71.2% F1 and 53.2% EM in MLQA, exceeding all strong baselines. A gain of 1.7/1.5% is obtained compared to the state-of-the-art X-MIXUP. As shown in Tables 2 and 3, our method also consistently outperforms all the strong baselines on XQuAD and TyDiQA. Our method obtains on average 1.8/2.2 and 2.6/3.5 improvement F1/EM scores compared to X-MIXUP. In conclusion, based on two backbones, our method outperforms state-of-the-art methods on three datasets, showing the effectiveness and generalization of our method. In addition, 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.

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

Table 2: Overall evaluation (F1/EM) over the TyDiQA dataset. * denotes the results of our re-implementation.

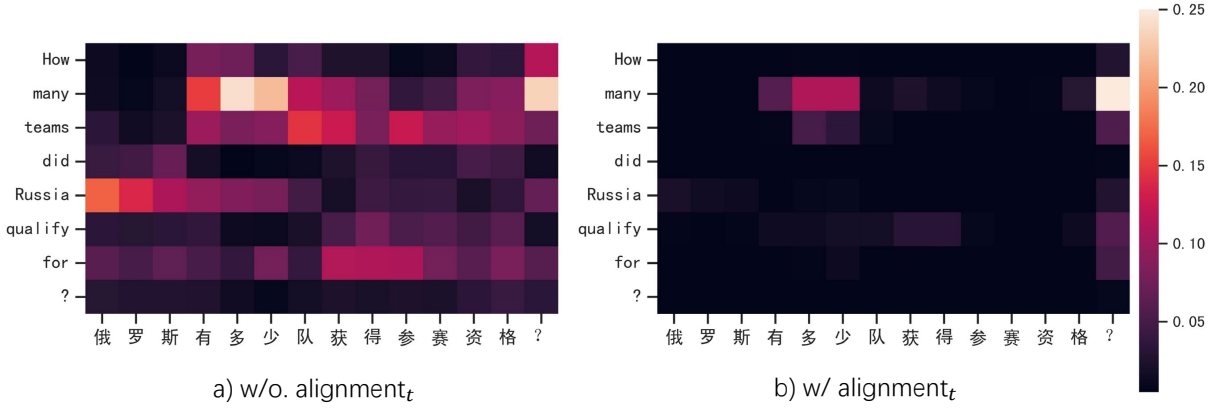

Figure 4: The attention distribution heat map of query part. We show the average result of multi-head attention.

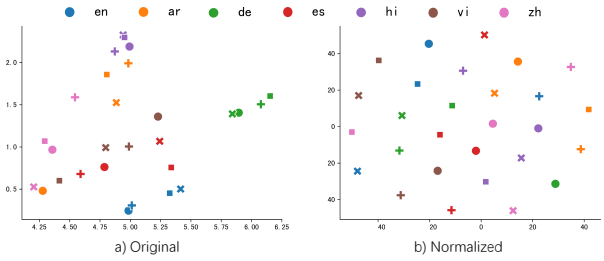

Figure 5: t-SNE distributions of sentence representations. Different shapes indicate different examples.

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 $\mathcal{L}_{tg}$. 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

Methods	en	ar	de	el	es	hi
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*	83.3/72.3	66.3/50.3	75.4/59.4	74.4/56.6	76.1/58.6	67.4/50.5
Trans-train*	83.9/73.0	71.1/54.9	78.1/62.9	76.7/59.4	80.2/62.3	75.0/59.7
AA-CL	84.1/73.1	66.5/50.3	77.9/62.5	-	80.0/61.7	73.8/58.9
X-MIXUP*	81.8/70.9	73.7/56.6	77.1/61.4	76.8/59.7	79.9/61.5	75.1/59.7
Ours	**86.0/74.9**	**77.7/61.2**	**81.4/66.1**	**80.3/63.6**	**82.7/65.4**	**79.4/62.9**
Methods	**ru**	**th**	**tr**	**vi**	**zh**	**Avg.**
Based on mBERT						
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
X-MIXUP	-	-	-	-	-	73.3/58.9
X-MIXUP*	80.1/63.9	61.9/55.7	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-R$_{base}$						
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	-	-	-	77.6/57.5	-	-
X-MIXUP*	77.8/62.1	72.7/67.1	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$_s$	71.2 / 53.0	78.5 / 64.0
w/o. alignment$_t$	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$_s$ and alignment$_t$ 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.

more distracted and not well aligned across languages. Instead, token-level alignment penalizes this behavior, allowing attention to be focused on the QA-related token (e.g., "many").

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-

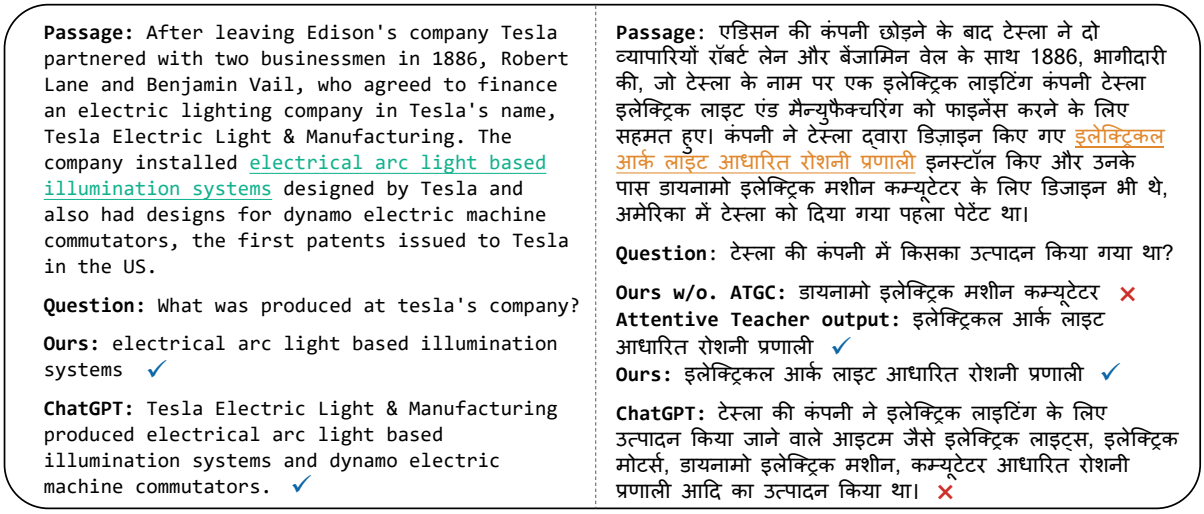

Figure 6: An example from XQuAD dataset, its ground-truth answer is marked with another color and underlined. The source language example (English) on the left corresponds to the low-resource target language (Hindi) example on the right. The ChatGPT used is Mar 14 Version, and our method uses mBERT as the backbone.

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

Prince Osei Aboagye, Yan Zheng, Chin-Chia Michael Yeh, Junpeng Wang, Wei Zhang, Liang Wang, Hao Yang, and Jeff M Phillips. 2022. Normalization of language embeddings for cross-lingual alignment.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2021. Multilingual transfer learning for qa using translation as data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12583–12591.

Nuo Chen, Linjun Shou, Ming Gong, and Jian Pei. 2022. From good to best: Two-stage training for cross-lingual machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10501–10508.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xilun Chen, Ahmed Hassan, Hany Hassan Awadalla, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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

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

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2021. Filter: An enhanced fusion method for cross-lingual language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12776–12784.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and ner. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1355–1360.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Shicheng Li, Pengcheng Yang, Fuli Luo, and Jun Xie. 2021. Multi-granularity contrasting for cross-lingual pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1708–1717.

Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, and Daxin Jiang. 2021. Calibrenet: Calibration networks for multilingual sequence labeling. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 842–850.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

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

OpenAI. 2023. ChatGPT. https://openai.com/chatgpt.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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.

Matīss Rikters and Mark Fishel. 2017. Confidence through attention. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 299–311.

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:

Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

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

Chengyu Wang, Minghui Qiu, Taolin Zhang, Tingting Liu, Lei Li, Jianing Wang, Ming Wang, Jun Huang, and Wei Lin. 2022. Easynlp: A comprehensive and easy-to-use toolkit for natural language processing. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 22–29.

Linjuan Wu, Shaojuan Wu, Xiaowang Zhang, Deyi Xiong, Shizhan Chen, Zhiqiang Zhuang, and Zhiyong Feng. 2022a. Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 991–1000.

Linjuan Wu, Shaojuan Wu, Xiaowang Zhang, Deyi Xiong, Shizhan Chen, Zhiqiang Zhuang, and Zhiyong Feng. 2022b. Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 991–1000.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, and Daxin Jiang. 2020. Enhancing answer boundary detection for multilingual machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 925–934.

Chen Zhang, Yuxuan Lai, Yansong Feng, Xingyu Shen, Haowei Du, and Dongyan Zhao. 2023. Cross-lingual question answering over knowledge base as reading comprehension. *arXiv preprint arXiv:2302.13241*.

Taolin Zhang, Junwei Dong, Jianing Wang, Chengyu Wang, Ang Wang, Yinghui Liu, Jun Huang, Yong Li, and Xiaofeng He. 2022. Revisiting and advancing chinese natural language understanding with accelerated heterogeneous knowledge pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 560–570. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *The 10th Conference on Lexical and Computational Semantics*, page 229.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417.

A Parameter Analysis

To further evaluate the effectivenss 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 γ achieving optimal performance on 0.1. We conjecture that when γ 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 is 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.

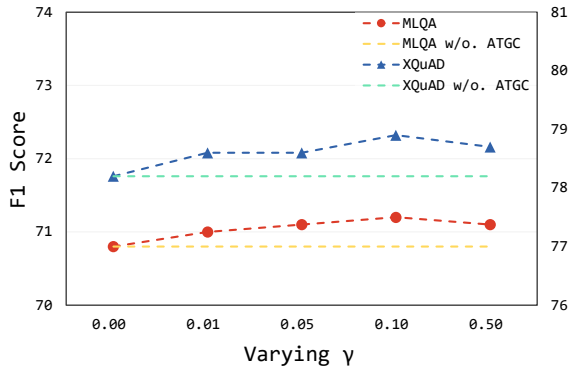

Figure 7: The parameter analysis of γ on MLQA (left vertical axis) and XQuAD (right vertical axis).

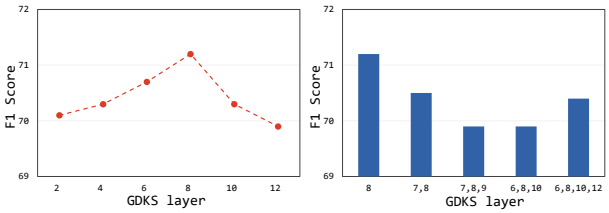

Figure 8: Performance comparison on implementing GDKS in different layers over the MLQA dataset.

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 transgressionally.

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. ✓

Passage: เมื่อยุโรปตะวันตกเฉียงเหนือค่อยๆ อุ่นขึ้นตั้งแต่ 22,000 ปีที่แล้ว เป็นต้นมา ชั้นใต้ผิวดินซึ่งกลายเป็นน้ำแข็งและธารน้ำแข็งหุบเขาซึ่งขยายตัวก็เริ่ม ละลาย และหิมะที่ตกในฤดูใบไม้ร่วงและฤดูหนาวก็ละลายในฤดูใบไม้ผลิ น้ำที่เกิดจากการละลายจำนวนมากไหลไปที่ แม่น้ำไรน์ และตามลำน้ำสาขาที่ไหลลง อากาศที่อุ่นขึ้นอย่างรวดเร็ว และการเปลี่ยนแปลงของพืชพันธุ์ในป่าเปิด เริ่มขึ้นเมื่อประมาณ 13,000 ก.ป. เมื่อถึงช่วง 9000 ปีก.ป. ยุโรปก็เต็มไปด้วยป่า เมื่อน้ำแข็งที่ปกคลุมทั่วโลกลดจำนวนลง ระดับน้ำในมหาสมุทรก็สูงขึ้น และช่องแคบอังกฤษกับทะเลเหนือก็โดนน้ำท่วมอีกครั้ง น้ำที่เกิดจากการละลายบวกกับน้ำในมหาสมุทรและการทรุดตัวของพื้นดิน ทำให้อดีตชายฝั่งของยุโรปจมน้ำ

Question: ยุโรปเต็มไปด้วยป่าและฟื้นฟูจากยุคน้ำแข็งสุดท้ายเมื่อใด

Ours: 9000 ปีก.ป. ✓

ChatGPT: ยุโรปเต็มไปด้วยป่าและฟื้นฟูจากยุคน้ำแข็งสุดท้ายเมื่อประมาณ 13,000 ก.ป. ✗

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. ✗

Passage: Bằng việc kết hợp định nghĩa của dòng điện giống như tốc độ thời gian thay đổi của điện tích, quy tắc nhân vectơ gọi là Định luật Lorentz mô tả lực tác dụng lên một điện tích chuyển động trong từ trường. Việc kết nối giữa điện và từ tính giúp mô tả một lực điện từ thống nhất tác dụng trên một điện tích. Lực này có thể được viết dưới dạng tổng của lực tĩnh điện (do điện trường) và lực từ (do từ trường). Nói đầy đủ, đây là định luật:

Question: Lực từ và lực điện nào tác dụng lên một điện tích?

Ours: lực điện từ thống nhất ✓

ChatGPT: Lực từ và lực tĩnh điện tác dụng lên một điện tích. ✗

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. ✓

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、国民服务、政府机构、英国商品和服务的共同市场、宪法、电力、煤炭、石油、燃气、核能、国防和国家安全、毒品政策、就业、对外政策和对欧关系、交通安全和管制的大多数方面、英国彩票、边境安全、社会保障以及英国的财政、经济和货币系统的稳定性。

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。 ✓

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