Toward Zero-Shot and Zero-Resource Multilingual Question Answering

CHIA-CHIH KUO AND KUAN-YU CHEN

Computer Science and Information Engineering Department, National Taiwan University of Science and Technology, Taipei 106, Taiwan

ABSTRACT

In recent years, multilingual question answering has been an emergent research topic and has attracted much attention. Although systems for English and other rich-resource languages that rely on various advanced deep learning-based techniques have been highly developed, most of them in low-resource languages are impractical due to data insufficiency. Accordingly, many studies have attempted to improve the performance of low-resource languages in a zero-shot or few-shot manner based on multilingual bidirectional encoder representations from transformers (mBERT) by transferring knowledge learned from rich-resource languages to low-resource languages. Most methods require either a large amount of unlabeled data or a small set of labeled data for low-resource languages. In Wikipedia, 169 languages have less than 10,000 articles, and 48 languages have less than 1,000 articles. This reason motivates us to conduct a zero-shot multilingual question answering task under a zero-resource scenario. Thus, this study proposes a framework to fine-tune the original mBERT using data from rich-resource languages, and the resulting model can be used for low-resource languages in a zero-shot and zero-resource manner. Compared to several baseline systems, which require millions of unlabeled data for low-resource languages, the performance of our proposed framework is not only highly comparative but is also better for languages used in training.

INDEX TERMS

Multilingual question answering, zero-shot, zero-resource, mBERT.

I. INTRODUCTION

With the increasing number of neural network architectures, such as recurrent neural networks and transformers [1], [2], various multilingual pretrained models have been developed and released [3], [4], [5], [6]. Among these models, the multilingual bidirectional encoder representations from transformers (mBERT) [3] is one of the most popular choices for downstream tasks due to its ease of use and state-of-the-art performances on various multilingual tasks. Although pre-training on large-scale multilingual corpora allows mBERT to be directly applied to the text of more than hundreds of different languages, inevitably, its performance in a low-resource language is generally worse than that in other rich-resource languages [7]. The scarcity of data makes it impractical to break the bottleneck merely by collecting more data. A simple but common solution is to utilize knowledge learned from a rich-resource language, such as English, to improve the performance of low-resource languages. Several studies have shown the potential of this approach through experiments in few-shot or zero-shot settings [8], [9], [10].

Because assistant applications are frequently installed in a variety of mobile phones and home devices around the world, multilingual question answering has been an emergent challenge in recent years. Furthermore, the global popularization of multimedia technology, video/audio sharing websites, and social networks has led to a significant growth in multilingual content nowadays. This has also increased the demand for machine reading comprehension of multilingual content, a typical case of multilingual question answering. In an example of machine reading comprehension, given a question and text passage, the machine is asked to predict a short text span from the text passage as the answer to the question.

Recent studies on cross-lingual transfer learning have focused on manipulating text representations to eliminate
language-specific information so that a shared latent space is created for all languages [11], [12], [13], [14], [15]. Ideally, language-agnostic text representations can be obtained to improve the effectiveness of transfer learning. Based on the analysis [15], the performance of downstream tasks increases as the Hausdorff distance between representations of two languages decreases. However, these studies rely either on a medium amount of labeled data or a large amount of unlabeled data, which are not always available for every low-resource language, considering that some languages only have a few hundred articles on Wikipedia. According to the official Wikipedia statistics, remarkably, 169 languages have less than 10,000 articles, and 48 languages only have less than 1,000 articles on Wikipedia. For example, Lingala and Hakka, as the mother tongues of 15 million and 50 million people, respectively, have only 3,226 and 9,416 articles on Wikipedia. Consequently, the scarcity of data can drastically decrease the practicality of a multilingual or monolingual system for a low-resource language. Moreover, most transfer learning techniques still require a medium to a large amount of unlabeled data, even in a zero-shot scenario.

Some recent studies have focused on zero-shot learning methods, which rely on a moderate amount of unlabeled data (more than 10,000 articles), but they are not always applicable to all languages due to data scarcity issues [12], [13]. Accordingly, this study aims at breaking the barrier and building a multilingual question answering framework for low-resource languages in a zero-shot and zero-resource manner. Formally, the proposed framework concentrates on fine-tuning mBERT toward a question answering system using labeled data from only rich-resource languages, and the resulting model can be directly deployed on low-resource languages with improved performance. Accordingly, it requires no labeled and unlabeled data for low-resource languages, making it a zero-shot and zero-resource multilingual QA model. In particular, in this study, all rich-resource and low-resource languages should be included in mBERT. We leave including languages that exclude mBERT as one of our emergent challenges in future works.

In sum, the major contributions of this study are at least threefold: First, a framework composed of four novel training strategies is introduced. The classic mBERT model can be fine-tuned using rich-resource language data based on the strategies, and the resulting model can be used to perform the multilingual question answering task for low-resource languages in a zero-shot and zero-resource manner while obtaining remarkable results. Second, to the best of our knowledge, the proposed framework is the first model that requires absolutely no unlabeled data under the zero-shot scenario. Third, in a series of experiments on the question answering task, the proposed framework achieves a competitive or better performance over baselines for nine highly diverse languages used in the training stage and five languages in a zero-shot scenario without using any unlabeled data.

II. RELATED WORKS
A. MULTILINGUAL PRETRAINING OF LANGUAGE REPRESENTATIONS

Before the multilingual variant of BERT was released, most studies focused on building multilingual word embeddings via unsupervised methods. Multilingual Unsupervised and Supervised Embeddings (MUSE) [31] is one of the most representative and powerful methods. It learns bilingual word mapping with relative criteria through multiple languages so that multiple sets of monolingual word embeddings can be smoothly merged into a set of multilingual word embeddings. However, even a robust method like MUSE still does not take any benefits from a large-scale multilingual corpus. As a result, the development of multilingual models is slowed down, causing a bottleneck in the performance.

Soon after the release of BERT went viral in the natural language processing (NLP) community, the multilingual variant of BERT was released in the next year. The multilingual BERT (mBERT) model easily outperforms MUSE on several multilingual datasets owing to its pretraining on a large-scale multilingual corpus crawled from Wikipedia, including articles of more than 100 languages. Afterward, a method named Language-Agnostic SEntence Representations (LASER) [32] was proposed, which forced word embeddings to be language-agnostic via training on machine translation tasks. Although the architecture of LASER is based on bidirectional long short-term memory, which is usually considered a weaker model compared to transformer-based architecture, LASER proved the importance of language agnosticism with its success.

The cross-lingual language model (XLM) [33] and its variant XLM-RoBERTa (XLM-R) [34] were thus proposed to combine the advantages of mBERT and LASER. Inspired by the masked language modeling task, a translation language modeling task was proposed for XLM-R. This task asks the model to perform classic masked language modeling (MLM) tasks with its input being the concatenation of a pair of parallel sentences of two languages. Owing to the knowledge learned in machine translation tasks and a pretraining corpus even larger than the one of mBERT, XLM-R easily outperforms mBERT on most multilingual datasets. However, while XLM-R adopts a SentencePiece [36] tokenizer, mBERT adopts a WordPiece tokenizer. Some researchers might prefer mBERT over XLM-R when developing a baseline system for word-level NLP tasks due to the extra technical difficulties of fixing word misalignment caused by the SentencePiece algorithm.

To further enhance the sentence-level language agnosticism for XLM-R, the information-theoretic framework for cross-lingual language model [37] proposed a cross-lingual contrastive learning method to maximize the mutual information between the representations of parallel sentences. Moreover, inspired by the adversarial approach of ELECTRA [38], XLM-E [39] manipulates the translation language modeling task of XLM-R into a task pretraining to ELECTRA style,
the translation replaced token detection (TRTD). In the TRTD task, the model first performs a classic translation language modeling prediction to recover the masked tokens in the parallel sentence input. Then, a discriminator model composed of a feed-forward fully connected neural network is employed to differentiate which tokens are recovered in the prediction.

**B. DOWNSTREAM FINE-TUNING OF LANGUAGE REPRESENTATIONS**

Although various pretraining methods for multilingual models swiftly push on the milestones of several multilingual benchmarks, the dreadfully high expense of pretraining a model on a large-scale multilingual corpus easily dissuades most researchers from developing a novel pretraining method for multilingual models. Instead, some researchers turn to analyzing the characteristics of these powerful pretrained models shown on different tasks so that a pretraining model can be even more powerful when properly applied to various downstream tasks. In a recent study, Rama et al. [16] analyzed the representations of mBERT using t-distributed stochastic neighbor embedding and found that the representations between two dissimilar languages have a sizable distributional difference, which could be detrimental to the effect of cross-lingual transfer learning. Thus, a modern multilingual NLP system usually performs worse in low-resource languages despite a large-scale corpus of rich-resource languages used for transfer learning. Accordingly, several studies have worked on improving the effectiveness of zero-shot cross-lingual transfer learning.

Liu et al. [13] proposed a zero-mean method, which directly reduces the distribution differences between representations of different languages for mBERT in a statistical approach. Formally, a mean vector can be obtained for a given language by mean-pooling all the hidden representations of all tokens in the training corpora. Then, the mean vector is subtracted from each original hidden representation for the training and inference stages of the downstream task. By doing so, language-specific information is expected to be removed, resulting in a better effect of cross-lingual transfer learning. The distributions of different languages can also be pulled together to easily transfer the learned knowledge to a zero-shot language.

Based on the zero-mean method, Liu et al. [13] further proposed a mean difference shift (MDS) method to transfer hidden representations from one language to another by manipulating the mean vectors of languages. English is the most rich-resource language in the pretraining of most multilingual models. Therefore, a rule of thumb is to transfer hidden representations of all languages for mBERT in a statistical way onto the distribution of a low-resource language. Meanwhile, as the generator model, mBERT is forced to output representations that can cheat the discriminator. As a result, the distributions of representations of different languages are all pulled toward English while fine-tuning mBERT so that the knowledge of other languages can be shared easily in the English domain.

Xia et al. [15] proposed MetaXL, a framework that utilizes meta-learning to train a representation transformation network layer to transform representations of a rich-resource language onto the distribution of a low-resource language. Then, the RTN layer transforms training examples of the rich-resource language before fine-tuning.

**III. METHODOLOGIES**

**A. VANILLA MULTILINGUAL BERT MODEL FOR QUESTION ANSWERING**

The application of multilingual BERT has been widely studied due to its state-of-the-art performance in several multilingual NLP-related tasks. The attributes of the transformer-based architecture make it a powerful encoder for learning knowledge from multiple languages all at once. When mBERT is applied in the multilingual question answering task, a naive approach is to employ it to encode a concatenation token (i.e., WordPiece) sequence of a passage and a question. Then, classification objectives are introduced to indicate a pair of start and end indices so that a text span can be extracted from the concatenation token sequence as the answer to the question. Formally, for a passage $p = \{w_1^p, w_2^p, \ldots, w_n^p\}$ and question $q = \{w_1^q, w_2^q, \ldots, w_n^q\}$, a concatenation token sequence $\{[CLS], w_1^q, \ldots, w_n^q, [SEP], w_1^p, \ldots, w_n^p, [SEP]\}$ can be obtained, where $[CLS]$ represents a special token of every concatenation token sequence and $[SEP]$ is a separator token. Next, the pretrained mBERT model extracts a set of hidden vectors for each token in the concatenation token sequence. Two fully connected feed-forward neural networks are adopted to individually translate the collection of hidden vectors to two sets of scores, corresponding to the start and end indices. As usual, a softmax function is then used to translate the scores to probability distributions. The training objective is computed and optimized to minimize the negative log-likelihood of the proper start and end indices for the concatenation token sequences for each training example:

$$\mathcal{L}_{QA} = - \log(P_{start_{gold}}^{p}) - \log(P_{end_{gold}}^{p}), \quad (1)$$

where $P_{start_{gold}}^{p}$ and $P_{end_{gold}}^{p}$ are the predicted probabilities of the ground-truth starting and ending positions, respectively.

**B. ZERO-MEAN AUGMENTATION**

This study aims to develop a zero-shot and zero-resource multilingual question answering framework that requires no labeled/unlabeled data for target low-resource languages. In other words, the deduced multilingual system is only constructed by rich-resource languages, and the resulting model can be directly deployed on low-resource languages with a
good performance. To reach the goal, a data augmentation method is first introduced.

As multiple languages appear in a multilingual dataset, we first collected a set of unlabeled text \( C_l \) for each language \( l \). Then, a token sequence \( W = \{[CLS], w_1, \ldots, w_n,[SEP]\} \in C_l \) was obtained for each text in \( C_l \) and encoded by the pretrained mBERT. Then, a language-specific mean vector \( \bar{v}_l \) was deduced by averaging all of the token representations generated by the \( k \)th transformer layer of the mBERT, where \( k \) is a manually adjusted hyperparameter. The process is depicted in Figure 1(a). Eventually, the mean vector \( \bar{v}_l \) is used only for fine-tuning the mBERT model.

C.-C. Kuo, K.-Y. Chen: Toward Zero-Shot and Zero-Resource Multilingual Question Answering

In particular, the mean vector is not needed for inference, so the resulting model can be used for any language in a zero-resource manner. Figure 1(b) provides a simple example of the workflow.

### C. Auxiliary Kullback–Leibler Divergence Objective

We propose an auxiliary training objective to further encourage mBERT for giving outputs of language-agnostic representations. Again, a concatenated token sequence of a training example was first encoded and augmented so that a pair of hidden outputs \( \{H^k, \bar{H}^k\} \) after the \( k \)th transformer layer is obtained. Then, \( \{H^k, \bar{H}^k\} \) were passed back to mBERT at its \((k+1)\)th transformer layer, leading to two pairs of probability distributions \( (p_{\text{start}}, p_{\text{end}}) \) and \( (\bar{p}_{\text{start}}, \bar{p}_{\text{end}}) \). Both \( p_{\text{start}} \) and \( p_{\text{end}} \) were derived from \( H \) and \( \bar{H} \) to denote the probability distributions of each token position being chosen as the start and end indices of the answer span, respectively. Similarly, \( \bar{p}_{\text{start}} \) and \( \bar{p}_{\text{end}} \) were obtained from \( \bar{H} \).

An auxiliary Kullback–Leibler divergence objective \( \mathcal{L}_{KL} \) was further introduced to guide the fine-tuning of mBERT:

\[
\mathcal{L}_{KL} = \frac{\text{KLDiv} (p_{\text{start}}, \bar{p}_{\text{start}}) + \text{KLDiv} (p_{\text{end}}, \bar{p}_{\text{end}})}{2},
\]

where \( p_{\text{start}} \) and \( p_{\text{end}} \) are used as references in the computation of Kullback–Leibler divergences.

The augmentation method and the auxiliary Kullback–Leibler divergence objective is adopted at the inference stage. In other words, the mean vectors are used only for fine-tuning the mBERT model, so the resulting model does not need a set of unlabeled data for obtaining a mean vector to a target language during inference. The idea is inspired and extended by Liu et al.’s zero-mean method [13], and a novel data augmentation method is introduced. In particular, the mean vector is not needed for inference, so the resulting model can be used for any language in a zero-resource manner. Figure 1(b) provides a simple example of the workflow.

**FIGURE 1.** (a) Computation of the mean vectors of mBERT for the data augmentation method. (b) Architecture of mBERT with the proposed data augmentation method and the auxiliary Kullback–Leibler divergence objective.
data augmentation method and auxiliary Kullback–Leibler divergence objective $\mathcal{L}_{KL}$.

D. INFORMATION COMPENSATION TRAINING STRATEGY

Inspired by the translation language modeling [34] task that concatenates the parallel text of two languages to learn the relationship between the two languages, we thus treated the original statistics and its zero-meaned statistics as a pair of parallel text. To make mBERT learn the relationship between the original statistics of a language and its zero-meaned statistics, we concatenated the original hidden states and zero-meaned hidden states as the input to the next transformer layer in mBERT so that mBERT can naturally learn the relationship between language-specific and language-agnostic information. Consequently, only the contextualized representations derived from the zero-meaned hidden states were retained and used to compute the auxiliary Kullback–Leibler divergence loss $\mathcal{L}_{KL}$ and classic question answering loss $\mathcal{L}_{QA}$. The method is named the information compensation training strategy, and Figure 2(a) shows the architecture of mBERT with the data augmentation method paired with the information compensation training strategy.

E. ZERO-MEANED TOKEN DETECTION

MLM is the core pretraining method of mBERT, which learns to recover randomly masked tokens in the input sequences. ELECTRA [38] is an adversarial pretraining method that further enhances the effectiveness of MLM by passing the recovering results of MLM to a discriminator model to detect whether a token is original or recovered by MLM so that MLM is forced to learn natural semantics. Inspired by ELECTRA, we propose a novel ELECTRA style training objective to further enhance the effectiveness of our framework. Formally, the contextual representations are taken from the last transformer layer in mBERT. Then, they are passed to a discriminator model, a layer of the feed-forward fully connected neural network, to detect whether a token is zero-meaned in the encoding process. We named the auxiliary objective zero-meaned token detection (ZMTD). Subsequently, for each contextual representation $h_i$ (or $\overline{h}_i$) and its corresponding label $y_i$, a binary classification training objective $\mathcal{L}_{ZMTD}$ is computed over all contextual representations:

$$\mathcal{L}_{ZMTD} = \sum_i \left( - (1 - y_i) \log P(0 | h_i) - y_i \log P(1 | \overline{h}_i) \right),$$

where $y_i$ is 1 if $\overline{h}_i$ is derived from a zero-meaned hidden state, and $\log P(1 | \overline{h}_i)$ represents the probability computed by the discriminator that $\overline{h}_i$ is derived from a zero-meaned hidden state. If the ZMTD is added to the model training, we randomly select 10% of input tokens to subtract the mean vector from their hidden states. Figure 2(b) illustrates the architecture of the proposed ZMTD method.

F. ENHANCED MULTILINGUAL BERT-BASED QUESTION ANSWERING (emBERTqa) MODEL

To create a zero-shot and zero-resource multilingual question answering model for low-resource languages, we introduce a data augmentation method, an auxiliary Kullback–Leibler divergence objective, an information compensation training strategy, and an auxiliary ZMTD objective. Empirically, a simple regularization is usually helpful in preventing overfitting, making the training processing stable, and achieving good performance for test data. In our implementation, we apply the L2 regularization on token embeddings at the beginning of the mBERT. Ultimately, an enhanced mBERT model, which combines all the methods by summing up their corresponding training objectives, is formulated:

$$\mathcal{L} = \mathcal{L}_{QA} + \mathcal{L}_{KL} + \mathcal{L}_{ZMTD} + \mathcal{L}_{L2}. \quad (5)$$

By leveraging all the proposed methods, we expect to obtain an enhanced multilingual BERT-based question answering (emBERTqa) model that can be used to answer questions in multiple languages. In particular, a major contribution of this work is that the proposed emBERTqa model is trained using rich-resource language data, and the resulting model can be used for low-resource languages without their unlabeled or labeled data. Consequently, to the best of our knowledge,
the proposed framework is the first modeling that requires absolutely no unlabeled data under the zero-shot scenario.

### IV. EXPERIMENTS

#### A. DATASETS AND SETUP

To examine the effectiveness of our proposed emBERTqa, we fine-tuned the mBERT model on the TyDiQA-GoldP dataset, which contains nine highly diverse languages from a typological perspective [17]. Then, we used the MLQA dataset as a zero-shot testing-only dataset [18]. Among all languages in the MLQA dataset, we report the scores of only five languages that are not covered in the TyDiQA-GoldP dataset, namely, German, Spanish, Chinese, Vietnamese, and Hindi. For the mean vectors we used in the data augmentation method, we collected the Wikipedia articles to compute the mean vectors following Liu’s zero-mean method [13], where more than five million tokens are involved in each language. The detailed statistics of the TyDiQA-GoldP and MLQA datasets are listed in Table 1.

We evaluated our framework using the average exact match (EM) and F1 scores. To evaluate the performance of target low-resource languages, we adopted the MACRO average of EM and F1 scores on the development set and testing set of the MLQA dataset. We applied the MACRO average scores as it can better demonstrate how well a multilingual model can evenly perform in each language.

We implemented our systems with the PyTorch deep learning library [40], Huggingface transformers library [41], and its provided pretrained parameters, the “bert-base-multilingual-uncased.” We used the AdamW algorithm and a learning rate of 3e-5 to optimize the model in the training stage. The batch size was set to 12 with a weight decay of 0.01 of the model parameters, and the gradients were clipped to 1.0 to prevent the model from overfitting.

#### B. BASELINE SYSTEMS

In the first set of experiments, several baselines were compared, including XLM-R [34], mBERT [3], adversarial learning method [12], MDS [13], and zero-mean method [13]. Table 2 summarizes the performances of these baselines on the MLQA dataset. XLM-R and mBERT are two classic multilingual pretrained language models, so we directly employed them for the question answering task without any fine-tuning. On the contrary, the adversarial learning method, MDS, and zero-mean method all require a large amount of unlabeled data for each low-resource language. Thus, XLM-R and mBERT perform multilingual question answering for low-resource languages in a zero-shot (i.e., without using labeled data) and zero-resource (i.e., without using unlabeled data) manner, while the other baseline systems belong to a zero-shot way. Although XLM-R generally performs better than mBERT in most sentence-level NLP tasks, such as text classification tasks, mBERT could perform better than XLM-R in word-level tasks, such as question answering, due to the difference between their tokenizers. The WordPiece tokenizer used by mBERT is based on a word-level tokenization algorithm, whereas the SentencePiece tokenizer used by XLM-R is based on a sentence-level algorithm, which might cause token misalignment when locating the start and end indices as the training labels of the question answering task (cf. Section II-A). Therefore, as shown in Table 2, XLM-R performs worse than mBERT in the MLQA dataset.

#### C. PROPOSED FRAMEWORK

We compared the proposed emBERTqa with the baseline models in the second set of experiments. All the results are listed in Table 2. The proposed emBERTqa generally outperformed the vanilla mBERT method and other baseline systems. Again, the adversarial learning method, MDS, and zero-mean method all require a large amount of unlabeled data for target languages in the MLQA dataset, whereas the proposed emBERTqa does not. Consequently, the major contribution of the proposed framework is that it requires neither labeled nor unlabeled data for low-resource languages. Moreover, the emBERTqa can even perform better than baseline models, which need unlabeled data. These results conclude that the proposed framework makes a step forward to create a set of language-agnostic representations and shared semantic space benefits from training data in various languages.

#### D. ABLATION STUDIES

Next, we evaluated different configurations of the proposed framework, and the results are shown in Table 2. Several observations can be drawn from the results. First, the data augmentation method can deliver better results than mBERT, demonstrating that the proposed method can provide performance gains and make the resulting model usable for multilingual question answering tasks in a zero-resource manner. Second, the auxiliary KL divergence objective helps the data augmentation method gain a slightly good performance, showing its beneficial effect of regularization in a zero-resource scenario. Third, the experimental results show that DataAug. + KLDiv. + L2 achieves better performances than DataAug. + KLDiv. This finding indicates that the simple regularization can indeed prevent overfitting, stabilize the

### TABLE 1. Statistics of the TyDiQA-GoldP and MLQA datasets.

| Language   | TyDiQA     | MLQA       |
|------------|------------|------------|
|            | Training   | Development| Testing   |
|            | 49,881     | 5,077      | 2,534     |
|            | 440        | 782        | 25,320    |
| English    | 3,696      | 921        |           |
| Arabic     | 2,390      | 113        |           |
| Bengali    | 6,855      | 565        |           |
| Finnish    | 5,702      | 276        |           |
| Indonisian | 6,490      | 499        |           |
| Telugu     | 5,563      | 669        |           |
| Spanish    | 500        | 5,253      |           |
| Germany    | 512        | 4,517      |           |
| Hindi      | 507        | 4,918      |           |
| Vietnamese | 511        | 5,495      |           |
| Chinese    | 504        | 5,137      |           |
TABLE 2. Zero-shot and zero-resource performance (in MACRO average) on the MLQA dataset.

| Method                  | Development | Zero-shot | Zero-shot + Zero-resource |
|-------------------------|-------------|-----------|---------------------------|
|                         | EM  | F1 | EM  | F1 | EM  | F1 |
| XLM-R                   | 34.19  | 50.99 | 35.07 | 52.24 |
| Adversarial Learning [12] | 35.21  | 51.81 | 35.14 | 51.84 |
| MDS [13]                | 35.62  | 52.17 | 35.86 | 52.83 |
| Zero-mean Method [13]   |          |         |        |      |
| emBERTqa                |          |         |        |      |
| DataAug                 |          |         |        |      |
| DataAug + KLDiv         |          |         |        |      |
| DataAug + KLDiv + L2    |          |         |        |      |
| DataAug + KLDiv + L2 + InformationCom |          |         |        |      |
| DataAug + KLDiv + L2 + ZMTD |          |         |        |      |

TABLE 3. Performance (in MICRO average) on the development set of the TyDiQA-GoldP dataset.

| Method                  | EM  | F1 |
|-------------------------|-----|----|
| mBERT                   | 67.96 | 78.64 |
| Adversarial Learning [12] | 67.42 | 79.63 |
| MDS [13]                | 67.89 | 79.39 |
| Zero-mean Method [13]   | 66.26 | 78.29 |
| emBERTqa                | 68.98 | 79.25 |

training process, and make the model robust. The information compensation training strategy is inspired by the translation language modeling task. Accordingly, we leveraged the idea to make the model learn the relationship between the original statistics of a language and its zero-meaned statistics. We also propose the ZMTD task to employ a discriminator to detect which representations of tokens are zero-meaned so that the model is forced to generate more language-agnostic representations. As shown in Table 2, the performance gains reveal that both methods provide positive benefits as expected.

E. IMPACT ON THE TRAINING LANGUAGES

In addition to the zero-shot and zero-resource scenario for low-resource languages, we also studied the potential impact of the approaches on rich-resource languages appearing in the fine-tuning stage. The comparison of the performance of the development set of the TyDiQA-GoldP dataset is presented in Table 3. Comparing the results in Tables 2 and 3, the zero-mean method as a baseline system has excellent performances for low-resource languages, but it sacrifices the performance of other languages used for fine-tuning, whereas our proposed emBERTqa can improve the performance of mBERT on both sides. Although our model cannot achieve the best results on the TyDiQA-GoldP dataset, it is still a very competitive method. In a nutshell, from Tables 2 and 3, we conclude that the proposed framework can be used to obtain a potential model (i.e., emBERTqa) that can be used in a zero-shot and zero-resource manner for target low-resource languages and can bring remarkable performances for rich- and low-resource languages.

V. CONCLUSION

This paper proposes a multilingual question answering framework, which requires neither labeled nor unlabeled data in a zero-shot and zero-resource scenario. When emBERTqa was evaluated on the MLQA and TyDiQA-GoldP datasets, it outperformed several baselines requiring a large amount of unlabeled data. A series of analyses demonstrated that better language-agnostic representations can be retrieved by emBERTqa to improve cross-lingual generalization capability. In summary, the proposed emBERTqa creates a potential way for low-resource languages on the multilingual question answering task. Hence, the framework can be generalized to other languages that exclude the original mBERT, and we leave its extension for future investigations. In the future, we also plan to leverage the framework for other NLP-related tasks, such as multilingual document retrieval and summarization.

ACKNOWLEDGMENT

The authors would like to thank the National Center for High-Performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

REFERENCES

[1] I. J. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017, arXiv:1706.03762.
[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., vol. 1, Jun. 2019, pp. 4171–4186.
[4] H. Huang, Y. Liang, N. Duan, M. Gong, L. Hou, D. Jiang, and M. Zhou, “Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks,” in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 2485–2494.
[5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in Proc. 58th Ann. Meeting Assoc. Comput. Linguistics, 2020, pp. 8440–8451.
[6] W. Qi, Y. Gong, Y. Yan, C. Xu, B. Yao, B. Zhou, B. Cheng, D. Jiang, J. Chen, R. Zhang, H. Li, and N. Duan, “ProphetNet-X: Large-scale pre-training models for English, Chinese, multi-lingual, dialog, and code generation,” 2021, arXiv:2104.08006.
[7] S. Wu and M. Dredze, “Are all languages created equal in multilingual BERT,” in Proc. 5th Workshop Represent. Learn. (NLP), 2020, pp. 120–130.
[8] A. Rahimi, Y. Li, and T. Cohn, “Massively multilingual transfer for NER,” in Proc. 57th Ann. Meeting Assoc. Comput. Linguistics, 2019, pp. 151–164.
