Multi-Lingual Question Generation with Language Agnostic Language Model

Bingning Wang, Ting Yao, Weipeng Chen, Jingfang Xu and Xiaochuan Wang
Sogou Inc.
Beijing, China
{wangbingning,yaoting}@sogou-inc.com

Abstract

Question generation is the task of generating coherent and relevant question given context paragraph. Recently, with the development of large-scale question answering datasets such as SQuAD, the English question generation has been rapidly developed. However, for other languages such as Chinese, the available training data is limited, which hinders the development of question generation in the corresponding language. To investigate the multi-lingual question generation, in this paper, we develop a language-agnostic language model, which learns the shared representation from several languages in a single architecture. We propose an adversarial training objective to encourage the model to learn both language-specific and language-independent information. We utilize abundant monolingual text to improve the multi-lingual question generation via pre-training. With the language-agnostic language model, we achieve significant improvement in multi-lingual question generation over five languages. In addition, we propose a large-scale Chinese question generation dataset containing more than 220k human-generated questions to benefit the multi-lingual question generation research.

1 Introduction

Question Generation (QG), also known as learning to ask, has attracted a lot of research interest in recent years. QG is regarded as the dual task of machine reading comprehension (Yuan et al., 2017; Xiao et al., 2018). Rather than answering a given question, learning to ask a coherent, relevant, and non-trivial question also requires a deep understanding of the context (Davey and McBride, 1986; Graesser et al., 2010), providing a good testbed for natural language understanding.

Conventional methods for question generation rely heavily on heuristic rules, and the standalone dependency parsing tool is needed to generate handcrafted templates (Mostow and Chen, 2009; Heilman and Smith, 2010; Rus et al., 2010; Hussein et al., 2014; Dhole and Manning, 2020). In recent years, with the development of deep learning and large-scale QA datasets, more and more neural network model has been proposed, which is also referred as neural question generation. Neural QG shows great advantage compared with previous rule-based systems in terms of both fluency and diversity of the generated questions (Duan et al., 2017; Yuan et al., 2017).

However, most progress in QG is made in English. For other languages such as Hindi, the lack of large-scale QG data limits its development. Recently, multi-lingual and cross-lingual language understanding has been studied in several NLP tasks, such as question answering (Liu et al., 2019; Cui et al., 2019), summarization (Zhu et al., 2019), natural language inference (Conneau et al., 2018), etc. For QG, Kumar et al. (2019) demonstrate that for low-resource Hindi, incorporating the large-scale English SQuAD (Rajpurkar et al., 2016) dataset could boost the QG result a lot.

For multi-lingual QG, a key factor is to learn a model that could transfer knowledge across different languages. In this paper, we propose a language-agnostic language model: it consists of the specific low-level module for each language, and a shared high-level module for multi-lingual information aggregation. Separating the language model into two levels enables us to learn the language-specific information in each language and the common information shared among languages. In this way, the knowledge in multi-lingual QG could be transferred via the high-level module.

For the language-agnostic language model, however, the distributed representation of the low-level module could be easily mixed with the language information, which makes the high-level module con-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2262–2272

tain some unnecessary language-specific features that are too specific to transfer across languages. Inspired by previous works on transfer learning (Chen et al., 2017; Liu et al., 2017), we propose an adversarial training objective to decouple the low-level module with the high-level module, which prevents the private and shared latent feature spaces from interfering with each other, making the high-level module language-invariant, thus achieving better transferability for different languages.

To get a better initialization for our model, we develop two self-supervised methods to pre-train our model on abundant monolingual text. We apply our model to five languages QG tasks that have human-labeled QG datasets. The experimental results demonstrate that all languages QG could benefit from the multi-lingual training. Our models surpass previous monolingual or multi-lingual QG methods by a large margin, even in zero-shot learning where we had no training data in the low-resource languages, our model achieves satisfactory results by merely trained on English dataset, which shows a promising transferability of the proposed model.

Besides, we also propose a large-scale Chinese QG dataset containing more than 220k human-labeled questions. We hope the proposed Chinese dataset could benefit the community for more comprehensive multi-lingual QG research. The codes and proposed datasets are available at https://github.com/benywon/LALM.

Our contributions are summarized as follow:

- We propose a novel language-agnostic language model which decouples the language specific and language independent information in QG.
- The proposed model achieves significant improvement over previous models in multi-lingual QG, and we analyze the transferability in multiple languages.
- We release a large-scale human labeled Chinese QG dataset containing more than 220k questions. To our best knowledge, this is the largest specific question generation dataset so far.

2 Related Work

Question generation has received increasing attention from the research community. Traditional QG systems are mostly rule-based, which sometimes utilizing off-the-shelf tools to get the syntactic structure, dependency relations, and semantic role of the passage (Mostow and Chen, 2009; Heilman and Smith, 2010). First, the target answers are generated using rules or semantic roles, next, low-quality questions are generated using hand-crafted rules or templates. Finally, the generated questions are ranked by features such as keyword matching degree or sentence perplexity (Hussein et al., 2014). The main drawbacks of these symbolic systems are that the rules and templates are expensive to manually create, and lack diversity.

With the development of deep learning and large-scale question answering datasets, motivated by neural machine translation, Du et al. (2017) proposed a sequence to sequence (seq2seq) architecture combined with attention mechanism, achieving a promising result on QA dataset SQuAD. Since then, many works have been proposed to extend the preliminary framework with rich features, such as named entity tags (Zhou et al., 2017) or answer position features (Duan et al., 2017), and incorporate copy mechanism to copy words from the context paragraph (Song et al., 2018). Other types of models are also introduced such as graph neural networks (Chen et al., 2019) or Transformer (Scialom et al., 2019). However, most of these works are focus on English QG and have not been validated in other languages.

Multi-Lingual language generation. Duan et al. (2019) translated documents as weakly supervised training data for zero-shot multi-lingual abstractive summarization. Chi et al. (2019) proposed a multi-lingual pre-training method that can transfer monolingual supervision signals to other pre-trained languages. Zhu et al. (2019) adopt large-scale supervised data from existing monolingual summarization datasets via translation strategy to perform multi-lingual summarization. Kumar et al. (2019) also proposed a multi-lingual question generation methods based on Transformer, they proposed a small Hindi QG dataset and improved the QG result on Hindi by training with additional English data.

Compared with the previous multi-lingual methods, our method directly separates the language-dependent module and language-independent module. We propose an adversarial decoupling module to improve the adaptive ability of the model. Besides, our model could be properly pre-trained by monolingual data, which obviates the need to construct the back-translation or pseudo-parallel data.

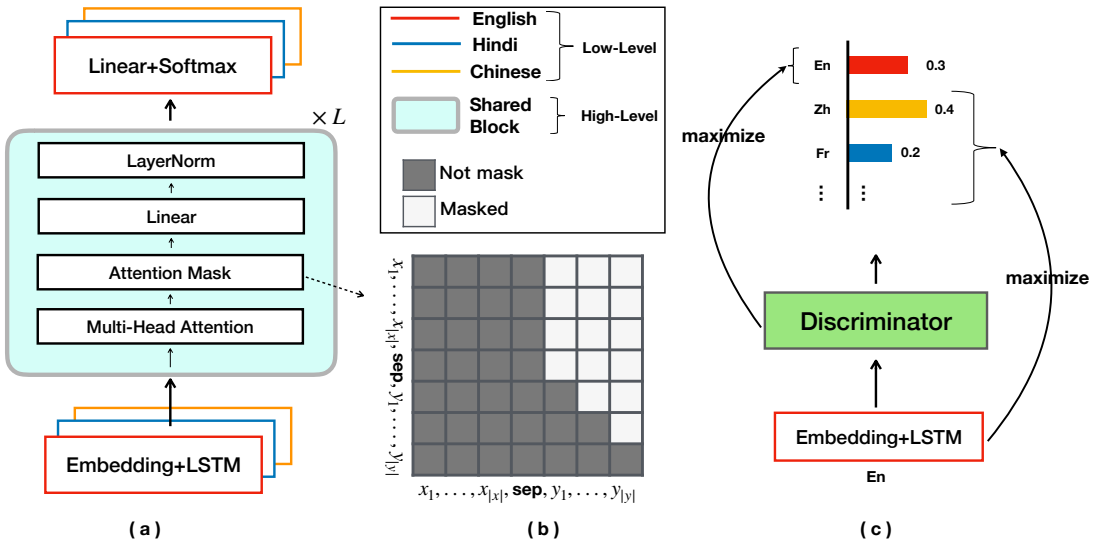

Figure 1: (a) The whole architecture of the proposed language-agnostic language model. It consists of the low-level language understanding module (Embedding+LSTM) and the high-level semantic understanding module (Transformer block), followed by a projection and softmax module. (b) The attention mask matrix M in the high-level module, $M_{i,j}$ means whether the word in position i could attend to the word in position j. The gray cells are allowed to attend and the others are masked to forbid attention. (c) The adversarial decoupling module where the discriminator tries to maximize the probability of the corresponding language while the generator (low-level module) tries to minimize it.

3 Language-agnostic Language Model

The Language-agnostic language model (LALM) consists of the low-level module and the high-level module, the whole architecture is illustrated in Figure 1(a) we describe it below.

3.1 Low-Level Module

The low-level module is built to perform the basic language understanding. In this paper, we adopt the LSTM (Hochreiter and Schmidhuber, 1997) encoder as the low-level language understanding module[1]. LSTM processes text in sequential order and embeds the language information into dense representations. We adopt the uni-directional LSTM in this paper to make the model auto-regressive.

For the language-agnostic language models, each language has its specific word embeddings and specific low-level language understanding LSTMs. This is different from some previous multi-lingual methods that a shared or aligned word embedding is utilized for different languages (Conneau et al., 2018; Lample and Conneau, 2019). Separating the language understanding module enables us to model specific linguistic characteristics in different languages. In Section 4, we will show that separating the low-level module for each language could benefit a lot for multi-lingual QG.

3.2 High-Level Module

The low-level module is built to perform the basic linguistic understanding, and the high-level module is built on top of the low-level module to perform higher-level information aggregation, which requires higher model capacity. In this paper, we use the Transformer (Vaswani et al., 2017) model as the high-level module.

The Transformer, with the core building-block called multi-head attention, has shown great advantages in representing languages in many NLP tasks. Current state-of-the-art models in natural language understanding benchmark GLEU[2] (Wang et al., 2018) are almost Transformer-based. In this paper, we focus on QG which is a sequence-to-sequence problem, so we adopt the mask operation similar with (Dong et al., 2019), which is illustrated in Figure 1(b). For a pair of sequence $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} = x_1, ..., x_{|x|}$ is the source, and $\mathbf{y} = y_1, ..., y_{|y|}$ is the target, we concatenate them together with a special token <sep>, forming a single sequence with length $|\mathbf{x}| + |\mathbf{y}| + 1$. We want all the positions in the source $\{1, 2, ..., |\mathbf{x}|\}$ to attend to each other so we can obtain the bi-directional representations

[1] In fact, we also conduct experiments on adopting other types of models as the low-level module such as Transformer or GRU, but the result is not comparable with the LSTM.

[2] https://gluebenchmark.com/

of the source, and all the positions in the target $\{|\mathbf{x}|+1, ..., |\mathbf{x}|+|\mathbf{y}|+1\}$ are forbidden to attend to future words:

$$M_{i,j} = \begin{cases} 0, & j \leq |\mathbf{x}| \quad \text{or} \quad j \leq i, \\ -\infty, & \text{otherwise.} \end{cases} \tag{1}$$

This attention mask operation enables us to build a causal language model that the generation of the current word only depends on its previous words. Therefore, the probability of $\mathbf{y}$ could be denoted as:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{|y|} p(y_i|y_{<i}, \mathbf{x}) \tag{2}$$

And the loss for the whole model is the negative log likelihood of the data:

$$\mathcal{L}_{NLL} = -\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y}} \sum \log p(y_i|y_{<i}, \mathbf{x}) \tag{3}$$

3.3 Adversarial Decoupling Module

In this paper, we want the representations of the low-level module in different languages to contain no language-specific information that is interleaved with the high-level module. In this way, the high-level module could focus on the semantic understanding shared across languages. We build a *discriminator* on top of the low-level module to determine whether the output of low-level representations contains the specific language information.

The discriminator is a bi-directional LSTM taking the output of the low-level module as input and tries to predict its language. Concretely, denote the output of the low-level module is $\mathbf{S} \in \mathbb{R}^{n,d}$ where n is the sequence length (i.e. $|\mathbf{x}| + |\mathbf{y}| + 1$), and d is the hidden size of the low-level module. The output of the discriminator can be represented as:

$$\begin{aligned} \mathbf{H} &= \text{bi-LSTM}(\mathbf{S}) \\ \mathbf{h} &= \text{Max-Pooling}(\mathbf{H}) \\ \hat{\mathbf{h}} &= \text{MLP}(\mathbf{h}) \\ \hat{y} &= \text{Softmax}(\hat{\mathbf{h}}) \end{aligned} \tag{4}$$

$\mathbf{h} \in \mathbb{R}^d$ is a pooled representations of the discriminator for classification. $\hat{y}$ is the language distribution in $\mathbb{R}^C$ where C is the number of languages. For the discriminator, the target is to maximize the probability of the corresponding language while the low-level module (generator) tries to minimize it. Therefore, they form an adversarial training objective that the low-level module must produce representations without discriminative language information. In this way, the discriminator acts as an adversarial decoupling module (ADM) to encourage the low-level module to generate language-agnostic representations.

The architecture of ADM is shown in Figure 1 (c), and the loss function for the discriminator and low-level module (generator) are:

$$\begin{aligned} \mathcal{L}_D &= -\log p(\hat{y}_i) \\ \mathcal{L}_G &= -\log p(1 - \hat{y}_i) \end{aligned} \tag{5}$$

where $\hat{y}_i$ is the discriminator probability for the input language i. In fact, the objective of the generator is to maximize the entropy of the discriminator's output to make it *less confident* of the language.

3.4 Pre-training

Recent works on NLP and language generation have shown the great advantage of large-scale pre-training (Devlin et al., 2018; Radford et al., 2019; Lewis et al., 2019; Roberts et al., 2020). In this paper, we also pre-train our model in massive multilingual text. Since our model is a sequence to sequence architecture, we develop two self-supervised objectives for language generation pre-training:

Denoised Auto-Encoder (DAE): Most previous works on natural language generation pre-training resort to DAE to initialize the model. In DAE, a corrupted version of the original sentence is created as the source and the model should reconstruct the original sentence. In this paper, we adopt the similar noising strategy as Lewis et al. (2019): (1) *Token Masking* random tokens are sampled and replaced with a special `[MASK]` token. (2) *Token Deletion* randomly deletes several tokens in the document. (3) *Token Replacement* randomly replace some tokens with other tokens in the vocabulary. (4) *Sentence Permutation* randomly swap some tokens in the sentence.

Next Sentence Generation: One of the problems of the DAE is that the input is always the corrupted sentence, which is not the case during fine-tune, the pretrain-finetune discrepancy may hurt the performance of the downstream tasks (Yang et al., 2019). Similar to Kiros et al. (2015) and Dong et al. (2019), we sample a consecutive segment in the text and divide it into two parts, we treat the first parts as the source and the second part as the target. The objective is to generate the second part based on the first part.

3.5 Question Generation Fine-tuning

After pre-training, we suppose the low-level module of our model has learned the multi-lingual linguistic information. Then the fine-tuning objective is to adjust the high-level module for question generation. Therefore, in this phase, we fix the low-level module, i.e. the word embedding, LSTM, and output projection linear layer, and only update the parameter of the high-level module.

4 Experiments

4.1 Dataset

The question generation datasets are sometimes directly derived from the corresponding question answering datasets. In the current question answering application, most multi-lingual datasets are automatically derived by translating from English SQuAD (Asai et al., 2018). However, it may reduce the multi-lingual QG tasks to translation tasks if we use these datasets. Therefore, we consider four different language QG datasets that are developed by the specific language speakers.

- **English (En)** We use the SQuAD (Rajpurkar et al., 2016) as the English question generation dataset. It is a standard machine reading comprehension data consists of nearly 100k human-labeled questions from Wikipedia.
- **Korean (Ko)** We use the Korquad1.0 (Lim et al., 2019) as the Korean QG data. It consists of more than 70,000 human-generated question-answer pairs on Korean Wikipedia articles.
- **French (Fr)** We adopt the French SQuAD-style dataset (d'Hoffschmidt et al., 2020) consisting of more than 25k human-curated French questions.
- **Hindi (Hi)** HiQuAD (Kumar et al., 2019) is a specific Hindi QG dataset containing 6,555 question-answer pairs. It was derived from the Hindi storybook.

Since the size of the QG dataset except English is comparative small, so we propose a new large-scale QG dataset created by humans on **Chinese (Zh)**. First of all, we collect nearly 3.5m passages from Baike[3], a Chinese Wikipedia-like encyclopedia. To increase the diversity of the selected paragraphs, we cluster the passages based on the bag-of-words, then we use Ward (Ward Jr, 1963) algorithm to select the centroid in each cluster, which result in nearly 100k passages. We ask volunteers to ask no more than 5 questions for each paragraph. Since we did not give the specific answer candidates for each paragraph, the annotators were encouraged to ask more general and comprehensive questions. We also ask other volunteers to check the quality and remove the questions that are either unanswerable or contain grammar errors. Finally, we obtain 224,962 question-paragraph pairs. We randomly select 180k of them as the training data, 20k samples for development, and the rest 24,962 for testing. We name it LAB (**L**earning to **A**sk on **B**aike).

	QG		Pre-train
	Train	Dev/Test	Name(Size)
En	86,635	8,965/8,964	enwiki(13.6Gb)
Zh	180,000	20,000/24,962	zhwiki(1.3Gb)
Ko	60,407	5,774/3,898	kowiki(608Mb)
Fr	20,731	3,188/2,189	frwiki(4.0Gb)
Hi	4,000	1,300/1,255	hiwiki(395Mb)

Table 1: The statistics of the multi-lingual pre-training data and question generation data.

We adopt the 2020-05-20 data dumps of the Wikipedia[4] in the corresponding language as the pre-training data. The details of the training data are shown in Table 1.

4.2 Implementation Details

In all experiments, we tokenize the text with sentencepiece (Kudo and Richardson, 2018). For all languages datasets, we set the vocabulary size to 30,000. We use the Adam (Kingma and Ba, 2014) optimizer with 5k warm-up steps and linearly decay the learning rate. $\beta_1, \beta_2, \epsilon$ was set to 0.9, 0.99 and 10^{-6}, respectively. For both pre-training and fine-tuning, the max learning rate was set to 10^{-4}. The batch size was 256 during pre-training and 64 during fine-tuning. We limit the max sequence length to 512. For the adversarial decoupling module training, following previous works of generative adversarial networks (Goodfellow et al., 2014; Salimans et al., 2016), the update rate for discriminator and generator was set to 1:10. For each of the 4 noising strategies in pre-training, we set the sample probability to 0.1. Similar with Scialom et al. (2019) we do not provide the answer and directly generate questions based on the context. We use three types of models:

LALM$_{\text{share}}$ is the shared language-agnostic lan-

[3] https://baike.sogou.com/

[4] https://dumps.wikimedia.org/

		Transformer	NQG++	Multi-BERT	CLQG	XNLG	$\text{LALM}_{\text{share}}$	$\text{LALM}_{\text{base}}$	$\text{LALM}_{\text{large}}$	$\text{LALM}_{\text{large}}$+ADM
En	BLEU-4	14.03	15.09	17.19	17.63	19.98	20.96	21.95	23.50	**24.94**
	METEOR	17.62	18.04	18.38	18.91	20.24	20.23	21.30	22.15	**23.28**
	ROUGE	40.79	40.24	44.82	43.34	46.51	47.47	48.23	50.34	**51.42**
Zh	BLEU-4	22.75	20.32	35.08	34.96	37.40	36.11	38.32	43.19	**44.10**
	METEOR	17.24	18.95	26.10	26.54	27.13	27.28	27.99	32.38	**33.04**
	ROUGE	30.14	29.87	38.46	40.11	42.15	43.25	44.49	45.16	**46.40**
Ko	BLEU-4	7.11	7.95	10.35	8.97	-	11.93	12.19	12.58	**12.93**
	METEOR	14.30	14.81	18.10	17.22	-	19.85	20.11	20.96	**21.10**
	ROUGE	22.17	24.13	31.28	29.34	-	34.10	34.88	34.79	**35.02**
Fr	BLEU-4	4.48	5.03	8.95	10.18	12.93	13.38	13.95	14.87	**15.28**
	METEOR	13.05	13.19	15.91	16.28	18.37	17.75	18.20	18.84	**19.92**
	ROUGE	32.17	31.66	39.34	41.23	40.96	41.15	42.80	43.11	**44.51**
Hi	BLEU-4	9.77	10.10	23.15	20.24	-	30.35	32.21	34.02	**35.19**
	METEOR	23.85	24.32	30.29	29.15	-	33.80	34.22	35.97	**36.25**
	ROUGE	33.16	34.91	41.06	40.64	-	48.82	49.14	50.94	**51.23**

Table 2: Main result of the multi-language QG. $\text{LALM}_{\text{share}}$ is similar with previous multi-lingual model that the parameters are shared across all languages. ADM represents the model trained with adversarial decoupling module.

guage model. It is similar with the proposed model but has no specific low-level LSTM for each language. That is, the low-level and high-level parameters are both shared across different languages. The hidden size was set to 768 and the layer size was set to 12, and each layer consists of 12 heads. We set the shared vocabulary size to 100,000.

$\text{LALM}_{\text{base}}$ is the base version of our model. It has the same hidden size as $\text{LALM}_{\text{share}}$. The low-level module was single layer uni-directional LSTM with hidden size 768. $\text{LALM}_{\text{base}}$ has nearly 138m parameters, where nearly half of them are low-level language understanding parameters.

$\text{LALM}_{\text{large}}$ is the large version of our proposed model. The hidden size, layer size, and head size were set to 1024,24,16, respectively. The low-level module was two-layer uni-directional LSTMs. $\text{LALM}_{\text{large}}$ has 548m parameters, where nearly a quarter of them are low-level module's parameters.

4.3 Criterion:

Following previous works of QG (Zhou et al., 2017; Chen et al., 2019), we adopt three widely used automatic metrics for evaluation: **BLEU**, **Meteor** and **Rouge-L**, which measure the n-gram similarities between the generated questions and real questions.

4.4 Baselines

We adopt 5 baseline methods for comparison.

- **Transformer** (Vaswani et al., 2017; Scialom et al., 2019) is the most widely used architecture in sequence-to-sequence learning. For each language, we train the correspondent Transformer model based on its training data. We set dropout ratio to 0.4 to prevent overfitting.
- **NQG++** (Zhou et al., 2017) is a popular neural QG model based on LSTM. It is enhanced with attention and copy mechanism[5].
- **Multi-BERT** (Devlin et al., 2018) is a multi-lingual extension to the original BERT model. It was trained on the multi-lingual wikipedia. All the language shares the same vocabulary. We adopt the way same with Rönnqvist et al. (2019) to extend BERT to language generation task.
- **CLQG** (Kumar et al., 2019) is a cross-lingual QG method based on Transformer. It is pre-trained by denoising autoencoders along with back-translation. We use the public implementation[6] and adopt the same word tokenization as well as pre-training data as our model.
- **XNLG** (Chi et al., 2019) is a multi-lingual language generation model that transfers monolingual supervision to all pre-trained languages. It was trained with English, Chinese and French datasets. We use their public pre-trained models[7] and fine-tune on the three QG dataset.

4.5 Multi-Lingual Question Generation

To evaluate the multi-lingual question generation ability of the proposed methods, we assemble all

[5] https://github.com/magic282/NQG
[6] https://github.com/vishwajeet93/clqg
[7] https://github.com/CZWin32768/XNLG

BLEU-4		
Zh		
P \ F	✓	✗
✓	38.32	36.03
✗	–	34.22
En		
P \ F	✓	✗
✓	21.95	20.61
✗	-	17.93

ROUGE		
Zh		
P \ F	✓	✗
✓	44.49	41.73
✗	–	40.12
En		
P \ F	✓	✗
✓	48.23	47.15
✗	-	45.02

Table 3: Multi-lingual and mono-lingual results for $\text{LALM}_{\text{base}}$. P denotes the pre-training and F denotes the fine-tuning, where ✓denote the multi-lingual while ✗denotes the mono-lingual training. For example, the upper right cell in each table denotes pre-training with multi-lingual but finetuning with mono-lingual.

QG data and train the LALM thereof. For Transformer and NQG++, we initialize the word embeddings by fasttext multilingual word embeddings (Grave et al., 2018). The result is shown in Table 2.

We can see from the table that our model excels at multi-lingual QG, achieving significant improvement over previous methods in all languages. Compared with other architectures such as Transformer, we explicitly separate the low-level and the high-level module in the proposed model and use adversarial networks to decouple them. Therefore, the shared high-level module is encouraged to learn more common representations across different languages, which is more transferable and benefits the downstream QG task a lot.

Besides, we can see that if we don't explicitly separate the low and high-level parameters ($\text{LALM}_{\text{share}}$), the result drops a lot. We hypothesis that different languages have different low-level language information, such as lexical, syntactical, etc. Embedding all language processing procedures into a single model may make the model hard to discriminate the language-specific information.

Besides, the model trained with the adversarial decoupling module achieves further improvement, the ADM may impose an implicit regularization on the low-level module to make the representations more abstract, and therefore encourage the high-level module to learn more common representations (Chen et al., 2017; Liu et al., 2017).

4.6 Human Evaluation

The automatic metrics are sometimes biased toward a specific attribute of the generated question (Hosking and Riedel, 2019). So we conduct human qualitative evaluation of the generated outputs. We consider three aspects of the generated questions: **Fluency**: Whether the generated questions are well-posed and natural, in terms of both grammar and semantic. **Answerable**: Whether the generated questions could be answered by the context paragraph. **Significance**: Whether the generated question is just a simple syntactical transformation of the paragraph sentence or trivial one that seems unlikely asked by human.

We randomly sample 50 generated questions from English and 50 from Chinese and ask three volunteers to evaluate the sample quality. The result is shown in Table 5. The result shows our proposed model is also excels at human evaluation, especially for significance, which is sometimes regarded as the most important factor in QG (Graesser et al., 2010). We also showcase some outputs of our model in Table 4. We can see that LALM could generate fluent and sound questions.

4.7 Multi-Lingual v.s. Mono-Lingual

Kumar et al. (2019) have found that in QG the performance of Hindi could be improved by training with additional English data. In this section, we evaluate whether the multi-lingual is superior to the mono-lingual QG. We focus on two aspects: **(1) Pre-training**. In contrast to the proposed multi-lingual pre-training, we adopt the mono-lingual pre-training where we only pretrain on specific languages[8] and fine-tune the QG models in the same language. **(2) Fine-tuning**. Different from the setup in Sec. 4.5 where we aggregate all languages QG data for training, we only fine-tune the model on specific language.

We experiment on English and Chinese with the LALM base model. The BLEU-4 and ROUGE-L scores are shown in Table 3. It is clear that for both pre-training and fine-tuning, the multi-lingual training improves the model a lot. Moreover, the multi-lingual plays a more important role in pre-training than in fine-tuning. We suppose that during pre-training, multiple languages perform a type of regularization on the shared high-level module, while in fine-tuning the language-dependent super-

[8]Therefore, we omit the adversarial decoupling module since it only takes effect on multi-lingual learning.

Table 4: Some generated cases of the proposed model.

English
Context: *The United Methodist Church opposes conscription as incompatible with the teaching of Scripture. Therefore, the Church supports and extends its ministry to those persons who conscientiously oppose all war, or any particular war, and who therefore refuse to serve in the armed forces or to cooperate with systems of military conscription. However, the United Methodist Church also supports and extends its ministry to those persons who conscientiously choose to serve in the armed forces or to accept alternative service. The church also states that "as Christians they are aware that neither the way of military action, nor the way of inaction is always righteous before God."*
Original: The Church supports those persons who conscientiously oppose what?
LALM: what does the church states after they oppose the construction ?
Chinese
Context: 电桥平衡#四个电阻*R0*、*R1*、*R2*、*Rx*连成四边形,称为电桥的四个臂。四边形的一个对角线连有检流计,称为"桥";四边形的另一对角线接上电源,称为电桥的"电源对角线"。*E*为线路中供电电源,学生实验用双路直流稳压电源,电压可在*0-30V*之间调节。*R*保护为较大的可变电阻,在电桥不平衡时取最大电阻作限流...
Original: 什么是电桥平衡?
LALM: 电桥平衡有什么用?
French
Context: *Le seul quartier d'habitation à avoir été fouillé est situé sur le site du Merkes, à l'est de la Voie processionnelle et du complexe sacré, entre les anciens quartiers de Ka-dingirra, Eridu et Shuanna. Sa voirie est caractérisée par des rues étroites approximativement rectilignes et se coupant quasiment à angles droits. Il s'agit peut-être de l'héritage d'un ancien plan orthogonal planifié qui a été altéré à la suite de remaniements de constructions, courants en raison de l'altération rapide des constructions en briques crues qui doivent régulièrement être restaurées.*
Original: En quoi sont fabriquées les habitations ?
LALM: Quelles sont les caractéristiques de la route ?
Korean
Context: 칭짱(藏)고원이라고도 불리는 티베트 고원은 동아시아에 위치한 넓고 높은 고원이다. 티베트자치구역과 중국 칭하이성(海省), 그리고 인도 카슈미르에 걸쳐 있는 티베트 고원은 남북 1000km, 동서 2500km에 뻗어 있으며, 그 평균 높이는 4500 미터가 넘는다. '세계의 지붕'으로 불릴 만큼 세계에서 가장 높고 크며 면적은 약 250만 평방 킬로미터나 된다. 이 고원은 인도-호주 플레이트와 유라시아 플레이트가 신생대에 충돌하며 생성되었으며 그 과정은 지금도 진행되고 있다. 이 고원은 산맥과 소금 호수가 분포한 고원의 건조 스텝지대를 형성하고 있다. 한 해 평균 강수량은 100mm에서 300mm로, 강수량의 대부분은 우박을 이룬다. 유목민들은 고원의 남부 및 동부 경계의 한 해 6개월가량 서리가 내리는 목초지에서 유목 생활을 유지하고 있다.
Original: 티베트고원의 면적은?
LALM: 티베트고원은어디에있습니까?

	Fluency	Answerable	Significance	Ave.
NQG++	1.01	1.09	1.02	1.04
Multi-BERT	1.22	1.19	1.23	1.21
$\text{LALM}_{\text{base}}$	**1.38**	**1.29**	**1.46**	**1.37**

Table 5: Human assessment of the generated questions on English and Chinese. Each question was assigned to score in $\{0,1,2\}$ which correspond to *bad*, *ok* and *excellent*, respectively. The result is statistical significant with $p < 0.05$.

vision of QG is more specific, which makes transfer learning less useful.

4.8 Zero-Shot Learning

In this section, we study the zero-shot multi-lingual learning ability of our model. The previous Section demonstrates that English SQuAD could strengthen other languages a lot. So we choose SQuAD as the training data and evaluate other languages. We only update the parameters of the high-level module for SQuAD without modifying the low-level language understanding module. Therefore, the replacement of the low-level module has little influence on the whole architecture, making the zero-shot inference available. We compare the zero-shot results of LALM base model with the supervised NQG++. The result is shown in Table 7.

We can see that the zero-shot version of our LALM appears to have equaled or eclipsed the QG ability of NQG++. It is an interesting result showing our model could transfer the question generation ability of English to other languages even without supervision. However, pure zero-shot learning is still struggle to achieve a good result, the supervision from the target language is necessary.

4.9 The Effect of Pre-training

We propose the self-supervised denoised auto-encoding and next sentence generation to pre-train the model. In this section, we construct a model that does not employ the pre-training but directly fine-tuned on the target data. The LALM hidden size to 256 and layer and head numbers of 4 and 8, respectively, to prevent overfitting. The results of English, Chinese, and Hindi are shown in Table 6. The performance of our model drops a lot without pre-training. Especially, it barely performs well for the low resource Hindi data because there

	En			Zh			Hi		
	B4	M	R	B4	M	R	B4	M	R
LALM	14.35	17.41	33.52	22.25	21.43	37.94	12.92	24.19	33.10
LALM+ADM	15.94	18.24	36.24	24.10	22.05	38.78	14.13	23.77	34.24
LALM+ADM+Pre-train	21.95	21.30	48.23	38.32	27.99	44.49	30.35	34.80	48.82

Table 6: Ablation study of the pre-training. The three models are fine-tuned on multi-lingual data.

		B1	B2	B3	B4	M	R
Zh	F	**43.07**	**31.04**	**23.58**	**17.74**	**18.06**	**22.44**
	Z	26.55	18.26	12.10	10.94	11.94	15.89
Kr	F	**25.20**	**15.34**	**10.71**	5.07	**14.35**	16.42
	Z	20.55	11.17	8.32	**5.95**	13.32	**16.77**
Fr	F	**31.31**	**14.91**	10.46	5.54	9.63	**23.02**
	Z	25.58	13.33	**12.49**	**6.32**	**11.06**	15.22
Hi	F	**30.15**	**20.42**	12.30	9.03	23.47	32.84
	Z	24.10	15.77	**12.54**	**10.89**	**26.42**	**33.01**

Table 7: Zero-shot multi-lingual evaluation. F denotes the performance of NQG++ model, and Z denotes zero-shot result where we fine-tune LALM base model on SQuAD and directly evaluate on other datasets.

are only 4,000 training instances. Nevertheless, when trained with the adversarial decoupling module, our model could achieve consistent improvement, demonstrating that the ADM is good at multi-lingual transfer learning.

5 Conclusion

In this paper, we propose a language-agnostic language model to deal with the multi-lingual question generation. The model consists of the low-level and the high-level module to explicitly represent the language-dependent and language-independent information, respectively. We operate the attention mask matrix to fit our model to the sequence to sequence learning. We propose an adversarial training mechanism to decouple the two-level modules, making the low-level module contains more abstractive representations and the high-level module language-agnostic. We also proposed a large-scale Chinese QG data containing more than 220k questions. Experiments on five languages demonstrate our model achieves significant improvements over previous methods in multi-lingual QG. For future work, we would like to apply our proposed model to other multi-lingual tasks such as summarization and question answering.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. And we appereate the dedicated labeling efforts contributed by the annoators, which makes the large-scale Chinese QG datasets avaliable for the community.

References

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275*.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuan-Jing Huang. 2017. Adversarial multi-criteria learning for chinese word segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1193–1203.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2019. Reinforcement learning based graph-to-sequence model for natural question generation. *arXiv preprint arXiv:1908.04942*.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2019. Cross-lingual natural language generation via pre-training. *arXiv preprint arXiv:1909.10481*.

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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. 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 1586–1595.

Beth Davey and Susan McBride. 1986. Effects of question-generation training on reading comprehension. *Journal of Educational Psychology*, 78(4):256.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé. 2020. Fquad: French question answering dataset. *arXiv preprint arXiv:2002.06071*.

Kaustubh D Dhole and Christopher D Manning. 2020. Syn-qg: Syntactic and shallow semantic rules for question generation. *arXiv preprint arXiv:2004.08694*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13042–13054.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. pages 1342–1352.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *EMNLP*, pages 866–874.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Art Graesser, Yasuhiro Ozuru, and Jeremiah Sullins. 2010. What is a good question?

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *HLT-NAACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Tom Hosking and Sebastian Riedel. 2019. Evaluating rewards for question generation models. In *NAACL*, pages 2278–2283.

Hafedh Hussein, Mohammed Elmogy, and Shawkat Guirguis. 2014. Automatic english question generation system based on template driven scheme. *IJCSI*, 11(6):45.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ICLR*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1. 0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. Xqa: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368.

Pengfei Liu, Xipeng Qiu, and Xuan-Jing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10.

Jack Mostow and Wei Chen. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *AIED*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

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.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual bert fluent in language generation? *arXiv preprint arXiv:1910.03806*.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2010. The first question generation shared task evaluation challenge.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6027–6032.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging context information for natural question generation. In *NAACL*, pages 569–574.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Han Xiao, Feng Wang, Yanjian Feng, and Jingyao Zheng. 2018. Dual ask-answer network for machine reading comprehension. *arXiv preprint arXiv:1809.01997*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Sandeep Subramanian, Saizheng Zhang, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. *arXiv preprint arXiv:1705.02012*.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *NLPCC*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.