Cross-Lingual Training for Automatic Question Generation

Vishwajeet Kumar[1,2], Nitish Joshi[2], Arijit Mukherjee[2], Ganesh Ramakrishnan[2], and Preethi Jyothi[2]

[1]IITB-Monash Research Academy, Mumbai, India
[2]IIT Bombay, Mumbai, India
{vishwajeet, nitishj, ganesh, pjyothi}@cse.iitb.ac.in
{arijitmukh007}@gmail.com

Abstract

Automatic question generation (QG) is a challenging problem in natural language understanding. QG systems are typically built assuming access to a large number of training instances where each instance is a question and its corresponding answer. For a new language, such training instances are hard to obtain making the QG problem even more challenging. Using this as our motivation, we study the reuse of an available large QG dataset in a secondary language (e.g. English) to learn a QG model for a primary language (e.g. Hindi) of interest. For the primary language, we assume access to a large amount of monolingual text but only a small QG dataset. We propose a cross-lingual QG model which uses the following training regime: (i) Unsupervised pretraining of language models in both primary and secondary languages and (ii) joint supervised training for QG in both languages. We demonstrate the efficacy of our proposed approach using two different primary languages, Hindi and Chinese. We also create and release a new question answering dataset for Hindi consisting of 6555 sentences.

1 Introduction

Automatic question generation from text is an important yet challenging problem especially when there is limited training data (i.e., pairs of sentences and corresponding questions). Standard sequence to sequence models for automatic question generation have been shown to perform reasonably well for languages like English, for which hundreds of thousands of training instances are available. However, training sets of this size are not available for most languages. Manually curating a dataset of comparable size for a new language will be tedious and expensive. Thus, it would be desirable to leverage existing question answering datasets to help build QG models for a

1. Sentence : विद्या के ये सभी रूप हमारे राष्ट्रीय ज्ञान के विविध अंग हैं
(All these forms of education are diverse aspects of our national knowledge system.)

Question (ground truth) : विद्या के सभी रूप हमारे राष्ट्रीय ज्ञान के क्या हैं ?
(What is the relationship between different forms of education and our national knowledge systems?)

Question (predicted) : विद्या के सभी रूप क्या हैं ?
(What are all the forms of education?)

2. Sentence : सभ्यता का अर्थ है संपत्ति की निरंतर वृद्धि , व्यवस्था और रक्षा अपनी संपत्ति की रक्षा औजारों के द्वारा की जाती है
(Civilization means continuous growth of prosperity, the system and its security are facilitated by the defense mechanism of the civilization.)

Question (ground truth) : सभ्यता का क्या अर्थ है ?
(What is the meaning of civilization?)

Question (predicted) : सभ्यता का क्या अर्थ है ?
(What is the meaning of civilization?)

Figure 1: Automatic QG from Hindi text.

new language. This is the overarching idea that motivates this work. In this paper, we present a cross-lingual model for leveraging a large question answering dataset in a secondary language (such as English) to train models for QG in a primary language (such as Hindi) with a significantly smaller question answering dataset.

We chose Hindi to be one of our primary languages. There is no established dataset available for Hindi that can be used to build question answering or question generation systems, making it an appropriate choice as a primary language. We create a new question answering dataset for Hindi (named **HiQuAD**): https://www.cse.iitb.ac.in/~ganesh/HiQuAD/clqg/. Figure 1 shows two examples of sentence-question pairs from **HiQuAD** along with the questions predicted by our best model. We also experimented with Chinese as a primary language. This choice was informed by our desire to use a language that was very different from Hindi. We use the same secondary language – English – with both choices of our primary language.

Drawing inspiration from recent work on unsupervised neural machine translation (Artetxe et al.,

2018; Yang et al., 2018), we propose a cross-lingual model to leverage resources available in a secondary language while learning to automatically generate questions from a primary language. We first train models for alignment between the primary and secondary languages in an unsupervised manner using monolingual text in both languages. We then use the relatively larger QG dataset in a secondary language to improve QG on the primary language. Our main contributions can be summarized as follows:

- We present a cross-lingual model that effectively exploits resources in a secondary language to improve QG for a primary language.

- We demonstrate the value of cross-lingual training for QG using two primary languages, Hindi and Chinese.

- We create a new question answering dataset for Hindi, **HiQuAD**.

2   Related Work

Prior work in QG from text can be classified into two broad categories.

Rule-based: Rule-based approaches (Heilman, 2011) mainly rely on manually curated rules for transforming a declarative sentence into an interrogative sentence. The quality of the questions generated using rule-based systems highly depends on the quality of the handcrafted rules. Manually curating a large number of rules for a new language is a tedious and challenging task. More recently, Zheng et al. (2018) propose a template-based technique to construct questions from Chinese text, where they rank generated questions using a neural model and select the top-ranked question as the final output.

Neural Network Based: Neural network based approaches do not rely on hand-crafted rules, but instead use an encoder-decoder architecture which can be trained in an end-to-end fashion to automatically generate questions from text. Several neural network based approaches (Du et al., 2017; Kumar et al., 2018a,b) have been proposed for automatic question generation from text. Du et al. (2017) propose a sequence to sequence model for automatic question generation from English text. Kumar et al. (2018a) use a rich set of linguistic features and encode pivotal answers predicted using a pointer network based model to automatically generate a question for the encoded
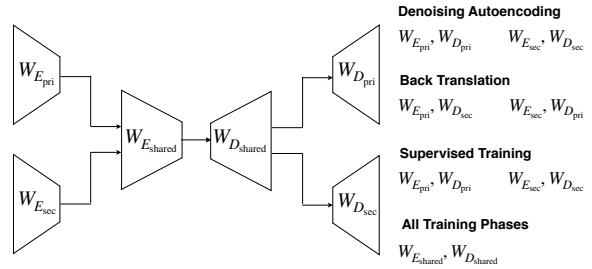

Figure 2: Schematic diagram of our cross-lingual QG system. $W_{E_{\mathrm{pri}}}$ and $W_{E_{\mathrm{sec}}}$ refer to parameters of the encoder layers specific to the primary and secondary languages; $W_{D_{\mathrm{pri}}}$ and $W_{D_{\mathrm{sec}}}$ are the weights of the corresponding decoder layers. $W_{E_{\mathrm{shared}}}$ and $W_{D_{\mathrm{shared}}}$ refer to weights of the encoder and decoder layers shared across both languages, respectively. Weights updated in each training phase are explicitly listed.

answer. All existing models optimize a cross-entropy based loss function, that suffers from exposure bias (Ranzato et al., 2016). Further, existing methods do not directly address the problem of handling important rare words and word repetition in QG. Kumar et al. (2018b) propose a reinforcement learning based framework which addresses the problem of exposure bias, word repetition and rare words. Tang et al. (2017) and Wang et al. (2017) propose a joint model to address QG and the question answering problem together.

All prior work on QG assumed access to a sufficiently large number of training instances for a language. We relax this assumption in our work as we only have access to a small question answering dataset in the primary language. We show how we can improve QG performance on the primary language by leveraging a larger question answering dataset in a secondary language. (Similarly in spirit, cross-lingual transfer learning based approaches have been recently proposed for other NLP tasks such as machine translation (Schuster et al., 2019; Lample and Conneau, 2019).)

3   Our Approach

We propose a shared encoder-decoder architecture that is trained in two phases. The first, is an **unsupervised pretraining** phase, consisting of denoising autoencoding and back-translation. This pretraining phase only requires sentences in both the primary and secondary languages. This is followed by a **supervised question generation** training phase that uses sentence-question pairs in both languages to fine-tune the pretrained weights.

1 Unsupervised Pretraining

while *not converged* **do**

2   Train autoencoder to generate sentence x_p from noisy sentence $\tilde{x}_p$ in primary language and similarly x_s from $\tilde{x}_s$ in the secondary language.

3   Back Translation: Generate sentences x_p' and x_s' in primary and secondary languages from x_s and x_p respectively, using the current translation model.

5   Train a new translation model using x_p' and x_s' where x_s and x_p are used for supervision, respectively.

end

6 Supervised Question Generation

7 Initialize with pretrained weights

while *not converged* **do**

8   Train sequence to sequence models for question generation in both the primary and secondary languages.

end

Algorithm 1: Cross-lingual Training Algorithm for QG

In Algorithm 1, we outline our training procedure and Figure 2 illustrates the overall architecture of our QG system. Our cross-lingual QG model consists of two encoders and two decoders specific to each language. We also enforce shared layers in both the encoder and the decoder whose weights are updated using data in both languages. (This weight sharing is discussed in more detail in Section 3.3.) For the encoder and decoder layers, we use the newly released Transformer (Vaswani et al., 2017) model that has shown great success compared to recurrent neural network-based models in neural machine translation. Encoders and decoders consist of a stack of four identical layers, of which two layers are independently trained and two are trained in a shared manner. Each layer of the transformer consists of a multi-headed self-attention model followed by a position-wise fully connected feed-forward network.

3.1 Unsupervised Pretraining

We use monolingual corpora available in the primary (Hindi/Chinese) and secondary (English) languages for unsupervised pretraining. Similar to Artetxe et al. (2018), we use denoising autoencoders along with back-translation (described in Section 3.1.1) for pretraining the language models in both the primary and secondary languages. Specifically, we first train the model to reconstruct their inputs, which will expose the model to the grammar and vocabulary specific to each language while enforcing a shared latent-space with the help

of the shared encoder and decoder layers. To prevent the model from simply learning to copy every word, we randomly permute the word order in the input sentences so that the model learns meaningful structure in the language. If x_p denotes the true input sentence to be generated from the sentence with permuted word order $\tilde{x}_p$ for the primary language, then during each pass of the autoencoder training we update the weights $W_{E_{\text{pri}}}$, $W_{E_{\text{shared}}}$, $W_{D_{\text{shared}}}$ and $W_{D_{\text{pri}}}$. For the secondary language, we analogously update $W_{E_{\text{sec}}}$, $W_{D_{\text{sec}}}$ and the weights in the shared layers as shown in Figure 2.

3.1.1 Back translation

In addition to denoising autoencoders, we utilize back-translation (Sennrich et al., 2016a). This further aids in enforcing the shared latent space assumption by generating a pseudo-parallel corpus (Imankulova et al., 2017).[1] Back translation has been demonstrated to be very important for unsupervised NMT (Yang et al., 2018; Lample et al., 2018). Given a sentence in the secondary language x_s, we generate a translated sentence in the primary language, $\tilde{x}_p$. We then use the translated sentence $\tilde{x}_p$ to generate the original x_s back, while updating the weights $W_{E_{\text{sec}}}$, $W_{E_{\text{shared}}}$, $W_{D_{\text{shared}}}$ and $W_{D_{\text{pri}}}$ as shown in Figure 2. Note that we utilize denoising autoencoding and back-translation for both languages in each step of training.

3.2 Supervised Question Generation

We formulate the QG problem as a sequence to sequence modeling task where the input is a sentence and the output is a semantically consistent, syntactically correct and relevant question in the same language that corresponds to the sentence. Each encoder receives a sentence $\mathbf{x}$ (from the corresponding language) as input and the decoder generates a question $\bar{\mathbf{y}}$ such that $\bar{\mathbf{y}} = \arg\max_y P(\mathbf{y}|\mathbf{x})$, and $P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{|y|} P(y_t|\mathbf{x}, y_{<t})$, where probability of each sub-word y_t is predicted conditioned on all the sub-words generated previously $y_{<t}$ and the input sentence $\mathbf{x}$. We initialize the encoder and decoder weights using unsupervised pretraining and fine-tune these weights further during the supervised

---

[1] A pseudo-parallel corpus consists of pairs of translated sentences using the current state of the model along with the original sentences.

QG model training. Specifically, in each step of training, we update the weights $W_{E_{\text{sec}}}$, $W_{E_{\text{shared}}}$, $W_{D_{\text{shared}}}$ and $W_{D_{\text{sec}}}$ using QG data in the secondary language and $W_{E_{\text{pri}}}$, $W_{E_{\text{shared}}}$, $W_{D_{\text{shared}}}$ and $W_{D_{\text{pri}}}$ using QG data in the primary language.

3.3 More Architectural Details

We make three important design choices:

1. **Use of positional masks:** Shen et al. (2018) point out that transformers are not capable of capturing within the attention, information about order of the sequence. Following Shen et al. (2018), we enable our encoders to use directional self attention so that temporal information is preserved. We use positional encodings which are essentially sine and cosine functions of different frequencies. More formally, positional encoding (PE) is defined as:

$$PE_{(\text{pos},2i)} = \sin\left(\frac{\text{pos}}{m^{\frac{2i}{d_{\text{model}}}}}\right) \quad (1)$$

$$PE_{(\text{pos},2i+1)} = \cos\left(\frac{\text{pos}}{m^{\frac{2i}{d_{\text{model}}}}}\right) \quad (2)$$

where m is a hyper-parameter, pos is the position, d_{model} is the dimensionality of the transformer and i is the dimension. Following Vaswani et al. (2017), we set m to 10000 in all our experiments. Directional self attention uses positional masks to inject temporal order information. Based on Shen et al. (2018), we define a forward positional mask (M^f) and a backward positional mask (M^b),

$$M_{ij}^f = \begin{cases} 0, & i < j. \\ -\infty, & \text{otherwise.} \end{cases}$$

$$M_{ij}^b = \begin{cases} 0, & i > j. \\ -\infty, & \text{otherwise.} \end{cases}$$

that processes the sequence in the forward and backward direction, respectively.

2. **Weight sharing:** Based on the assumption that sentences and questions in two languages are similar in some latent space, in order to get a shared language independent representation, we share the last few layers of the encoder and the first few layers of the decoder (Yang et al., 2018). Unlike Artetxe et al. (2018); Lample et al. (2018), we do not share the encoder completely across the two languages, thus allowing the encoder layers private to each language to capture language-specific information. We found this to be useful in our experiments.

3. **Subword embeddings**: We represent data using BPE (Byte Pair Encoding) (Gage, 1994) embeddings. We use BPE embeddings for both unsupervised pretraining as well as the supervised QG training phase. This allows for more fine-grained control over input embeddings compared to word-level embeddings (Sennrich et al., 2016b). This also has the advantage of maintaining a relatively smaller vocabulary size.[2]

4 Experimental Setup

We first describe all the datasets we used in our experiments, starting with a detailed description of our new Hindi question answering dataset, "**HiQuAD**". We will then describe various implementation-specific details relevant to training our models. We conclude this section with a description of our evaluation methods.

4.1 Datasets

4.1.1 HiQuAD

HiQuAD (Hindi Question Answering dataset) is a new question answering dataset in Hindi that we developed for this work. This dataset contains 6555 question-answer pairs from 1334 paragraphs in a series of books called Dharampal Books. [3]

Similar to SQuAD (Rajpurkar et al., 2016), an English question answering dataset that we describe further in Section 4.1.2, HiQuAD also consists of a paragraph, a list of questions answerable from the paragraph and answers to those questions. To construct sentence-question pairs, for a given question, we identified the first word of the answer in the paragraph and extracted the corresponding sentence to be paired along with the question. We curated a total of 6555 sentence-question pairs.

We tokenize the sentence-question pairs to remove any extra white spaces. For our experiments, we randomly split the HiQuAD dataset into train,

---

[2]Using word embeddings across pretraining and the main QG task makes the vocabulary very large, thus leading to large memory issues.

[3]HiQuAD can be downloaded from: `https://www.cse.iitb.ac.in/~ganesh/HiQuAD/clqg/`

#pairs (Train set)	4000
#pairs (Dev set)	1300
#pairs (Test set)	1255
Text: avg tokens	28.64
Question: avg tokens	14.13

Table 1: HiQuAD dataset details

development and test sets as shown in Table 1. All model hyperparameters are optimized using the development set and all results are reported on the test set.

4.1.2 Other Datasets

We briefly describe all the remaining datasets used in our experiments. (The relevant primary or secondary language is mentioned in parenthesis, alongside the name of the datasets.)

IITB Hindi Monolingual Corpus (Primary language: **Hindi**) We extracted 93,000 sentences from the IITB Hindi monolingual corpus[4] , where each sentence has between 4 and 25 tokens. These sentences were used for unsupervised pretraining.

IITB Parallel Corpus (Primary language: **Hindi**) We selected 100,000 English-Hindi sentence pairs from IITB parallel corpus (Kunchukuttan et al., 2018) where the number of tokens in the sentence was greater than 10 for both languages. We used this dataset to further fine-tune the weights of the encoder and decoder layers after unsupervised pretraining.

DuReader (He et al., 2018) **Chinese Dataset:** (Primary language: **Chinese**) This dataset consists of question-answer pairs along with the question type. We preprocessed and used "DESCRIPTION" type questions for our experiments, resulting in a total of 8000 instances. From this subset, we created a 6000/1000/1000 split to construct train, development and test sets for our experiments. We also preprocessed and randomly extracted 100,000 descriptions to be used as a Chinese monolingual corpus for the unsupervised pretraining stage.

News Commentary Dataset: (Primary language: **Chinese**) This is a parallel corpus of

---

news commentaries provided by WMT.[5] It contains roughly 91000 English sentences along with their Chinese translations. We preprocessed this dataset and used this parallel data for fine-tuning the weights of the encoder and decoder layers after unsupervised pretraining.

SQuAD Dataset: (Secondary language: **English**) This is a very popular English question answering dataset (Rajpurkar et al., 2016). We used the train split of the pre-processed QG data released by Du et al. (2017) for supervised QG training. This dataset consists of 70,484 sentence-question pairs in English.

4.2 Implementation Details

We implemented our model in TensorFlow.[6] We used 300 hidden units for each layer of the transformer with the number of attention heads set to 6. We set the size of BPE embeddings to 300. Our best model uses two independent encoder and decoder layers for both languages, and two shared encoder and decoder layers each. We used a residual dropout set to 0.2 to prevent overfitting. During both the unsupervised pretraining and supervised QG training stages, we used the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $1e-5$ and batch size of 64.

4.2.1 Unsupervised Pretraining

For Hindi as the primary language, we use 93000 Hindi sentences from the IITB Hindi Monolingual Corpus and around 70000 English sentences from the preprocessed SQuAD dataset for unsupervised pretraining. We pretrain the denoising autoencoders over 15 epochs. For Chinese, we use 100000 Chinese sentences from the DuReader dataset for this stage of training.

4.2.2 Supervised Question Generation Training

We used 73000 sentence-question pairs from SQuAD and 4000 sentence-question pairs from HiQuAD (described in Section 4.1.1) to train the supervised QG model in Hindi. We used 6000 Chinese sentence-question pairs from the DuReader dataset to train the supervised QG model in Chinese. We initialize all the weights, including the BPE embeddings, from the pretraining phase and fine-tune them until convergence.

Language	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Hindi	Transformer	28.414	18.493	12.356	8.644	23.803	29.893
	Transformer+pretraining	41.059	29.294	21.403	16.047	28.159	39.395
	CLQG	41.034	29.792	22.038	16.598	27.581	39.852
	CLQG+parallel	**42.281**	**32.074**	**25.182**	**20.242**	**29.143**	**40.643**
Chinese	Transformer	25.52	9.22	5.14	3.25	7.64	27.40
	Transformer+pretraining	30.38	14.01	8.37	5.18	**10.46**	**32.71**
	CLQG	**30.69**	**14.51**	**8.82**	5.39	10.44	31.82
	CLQG+parallel	30.30	13.93	8.43	**5.51**	10.26	31.58

Table 2: BLEU, METEOR and ROUGE-L scores on the test set for Hindi and Chinese question generation. Best results for each metric (column) are highlighted in **bold**.

4.3 Evaluation Methods

We evaluate our systems and report results on widely used BLEU (Papineni et al., 2002), ROUGE-L and METEOR metrics. We also performed a human evaluation study to evaluate the quality of the questions generated. Following Kumar et al. (2018a), we measure the quality of questions in terms of syntactic correctness, semantic correctness and relevance. Syntactic correctness measures the grammatical correctness of a generated question, semantic correctness measures naturalness of the question, and relevance measures how relevant the question is to the text and answerability of the question from the sentence.

5 Results

We present our automatic evaluation results in Table 2, where the primary language is Hindi or Chinese and the secondary language in either setting is English. We do not report on Chinese as a secondary language owing to the relatively poor quality of the Chinese dataset. Here are all the models we compare and evaluate:

• **Transformer**: We train a transformer model (Vaswani et al., 2017) using the QG dataset in the primary language. This serves as a natural baseline for comparison.[7] This model consists of a two-layer encoder and a two-layer decoder.

• **Transformer+pretraining**: The above-mentioned **Transformer** model undergoes an additional step of pretraining. The encoder and decoder layers are pretrained using monolingual data from the primary language. This model will help further demonstrate the value of cross-lingual training.

• **CLQG**: This is our main cross-lingual question generation model (described in Section 3)

---

[7] We also trained a sequence-to-sequence model by augmenting HiQuAD with SQuAD sentences translated into Hindi using Google Translate. This did not perform well giving a BLEU-4 score of 7.54.

where the encoder and decoder layers are initialized in an unsupervised pretraining phase using primary and secondary language monolingual corpora, followed by a joint supervised QG training using QG datasets in the primary and secondary languages.

• **CLQG+parallel**: The CLQG model undergoes further training using a parallel corpus (with primary language as source and secondary language as target). After unsupervised pretraining, the encoder and decoder weights are fine-tuned using the parallel corpus. This fine-tuning further refines the language models for both languages and helps enforce the shared latent space across both languages.

We observe in Table 2 that **CLQG+parallel** outperforms all the other models for Hindi. For Chinese, parallel fine-tuning does not give significant improvements over CLQG; this could be attributed to the parallel corpus being smaller in size (when compared to Hindi) and domain-specific (i.e. the news domain).

6 Discussion and Analysis

We closely inspect our cross-lingual training paradigm using (i) a human evaluation study in Section 6.1 (ii) detailed error analysis in Section 6.2 and (iii) ablation studies in Section 6.3. All the models analyzed in this section used Hindi

Model	Syntax		Semantics		Relevance	
	Score	Kappa	Score	Kappa	Score	Kappa
Transformer	71	0.239	62.5	0.46	32	0.75
CLQG +parallel	**72**	0.62	**68.5**	0.82	**54**	0.42

Table 3: Human evaluation results as well as inter-rater agreement (column "Kappa") for each model on the Hindi test set. The scores are between 0-100, 0 being the worst and 100 being the best. Best results for each metric (column) are in **bold**. The three evaluation criteria are: (1) syntactic correctness (Syntax), (2) semantic correctness (Semantics), and (3) relevance to the paragraph (Relevance).

Sentence : आज देश में जो हो रहा है वह तो एक बहुत निचले स्तर का यूरोप व अमरीका का अनुकरण हो रहा है
(What is happening in the country today is a very low level emulation of Europe and America.)

Question (human generated) : आज भारत देश में जो हो रहा है वह है ?
(How do you describe whatever is happening in India today?)

Question (predicted) : आज भारत देश में जो कुछ हो रहा है वह क्या है ?
(How do you describe whatever is happening in India today?)

(a)

Sentence : प्लेफेयर ने कहा कि गुरुत्वाकर्षण सिद्धांत एवं इण्टीग्रल केलकुलस के गणितीय सिद्धांतों के ज्ञान के बिना भारतीय गणितज्ञ इतना अचूक गणित ज्योतिषीय आकलन कर ही नहीं सकते थे
(Playfair said that without the knowledge of the mathematical principles of ....)

Question (human generated) : प्लेफेयर ने क्या कहा ?
(What did Playfair say?)

Question (predicted) : प्लेफेयर ने क्या कहा ?
(What did Playfair say?)

(b)

Sentence : इस गाथा के अनुसार ब्रह्म के तप व संकल्प से सृष्टि का सर्जन होता है , और फिर यह अनेकानेक आवर्तनों से होती हुई , वापस ब्रह्म में लीन हो जाती है
(According to this narrative, the universe is created by tenacity and resolution of...)

Question (human generated) : इस गाथा के अनुसार किससे सृष्टि का सर्जन होता है ?
(According to this narrative, how is the universe created?)

Question (predicted) : किस चीज़ के अनुसार सृष्टि का सर्जन होता है ?
(According to what the universe is created?)

(c)

Figure 3: Three examples of correctly generated Hindi questions by our model, further analyzed in Section 6.2.

Sentence : इसी ईसाईकरण का दूसरा नाम पश्चिमीकरण है , जिसे करने के प्रयत्न स्वतंत्र भारत की सरकारें भी करती चली आ रही हैं
(The second name of this Christianization is Westernization, which independent India's governments has been trying to do.)

Question (human generated) : ईसाईकरण का दूसरा नाम क्या है ?
(What is the second name of Christianization?)

Question (predicted) : विज्ञान का दूसरा नाम क्या है ?
(What is the second name of science?)

(a)

Sentence : हम जानते हैं कि अरब बहुत बड़ा विदेश व्यापार करते थे
(We know that the Arabs used to very big foreign trade.)

Question (human generated) : अरब क्या करते थे ?
(What did Arab people used to do?)

Question (predicted) : अरब लोग किस तरह के थे ?
(What kind of people were the Arabs?)

(b)

Figure 4: Two examples of incorrectly generated Hindi questions by our model, further analyzed in Section 6.2.

Sentence : 打开 微信 ， 点击 " 我 " ， 选择 通用 ， 点击 功能 ， 选择 群发 助手 ， 点 开始 群发 ， 如果 被 对方 删了 发布 出去.
(Open WeChat, click "I", select General, click on function, select the group assistant, click to start the group, if it is deleted by the other party, release it.)

Question (human generated) : 怎么 知道 对方 微信 是否 把 我 删 了 ?
(How do I know if I have been deleted by the other person's Wechat?)

Question (predicted) : 怎样 知道 微信 好友 是否 删除 自己 ？
(How do I know if my WeChat friends deleted me?)

(a)

Sentence : 放置 在 冰箱 里 ； 把 百香果 洗干净 切成 条 放在 太阳 底下 晒 成果 干.
(Put them in the refrigerator; wash and cut them into strips and dry them in the sun.)

Question (human generated) : 百香果 怎么 保存 得 久 一点 ?
(How can fruit be stored for longer ?))

Question (predicted) : 樱桃 怎么 保存 ？
(How to store cherries?)

(b)

Figure 5: Automatic QG from Chinese text.

as the primary language.[8]

6.1 Human evaluation

We conduct a human evaluation study comparing the questions generated by the **Transformer** and **CLQG+parallel** models. We randomly selected a subset of 100 sentences from the Hindi test set and generated questions using both models. We presented these sentence-question pairs for each model to three language experts and asked for a binary response on three quality parameters namely syntactic correctness, semantic correctness and relevance. The responses from all the experts for each parameter was averaged for each model to get the final numbers shown in Table 3. Although we perform comparably to the baseline model on syntactic correctness scores, we obtain significantly higher agreement across annotators using our cross-lingual model. Our cross-lingual model performs significantly better than

the **Transformer** model on "Relevance" at the cost of agreement. On semantic correctness, we perform signficantly better both in terms of the score and agreement statistics.

6.2 Error Analysis

Correct examples: We show several examples where our model is able to generate semantically and syntactically correct questions in Figure 3. Figure 3b shows our model is able to generate questions that are identical to human-generated questions. Fig. 3c demonstrates that our model can generate new questions which clearly differ from the human-generated questions but are syntactically correct, semantically correct and relevant to the text. Fig. 3a shows a third question which differs from the human-generated question in only a single word but does not alter its quality.

Incorrect examples: We also present a couple of examples where our model is unable to generate good questions and analyze possible reasons for the same. In Fig. 4a, the model captures the type of

---

[8]Figure 5 shows two examples of correctly generated Chinese questions.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
CLQG (no pretraining)	31.707	20.727	13.954	9.862	24.209	32.332
CLQG	41.034	29.792	22.038	16.598	27.581	39.852
CLQG+ parallel	**42.281**	**32.074**	**25.182**	**20.242**	**29.143**	**40.643**

Table 4: Ablation study showing the importance of both unsupervised and unsupervised pretraining for Hindi

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Hindi QG only	41.66	31.576	24.572	19.538	28.665	**40.765**
Hindi QG + English QG	**42.281**	**32.074**	**25.182**	**20.242**	**29.143**	40.643

Table 5: Ablation study showing the importance of using English QG data for Hindi QG

question correctly but gets the main subject of the sentence wrong. On the other hand, Fig. 4b shows a question which is syntactically correct and relevant to the main subject, but is not consistent with the given sentence.

6.3 Ablation Studies

We performed two experiments to better understand the role of each component in our model towards automatic QG from Hindi text.

6.3.1 Importance of unsupervised pretraining

We construct a model which does not employ any unsupervised or supervised pretraining but uses the same network architecture. This helps in studying the importance of pretraining in our model. We present our results in Table 4. We observe that our shared architecture does not directly benefit from the English QG dataset with simple weight sharing. Unsupervised pretraining (with back-translation) helps the shared encoder and decoder layers capture higher-level language-independent information giving an improvement of approximately 7 in BLEU-4 scores. Additionally, the use of parallel data for fine-tuning unsupervised pretraining aids this process further by improving BLEU-4 scores by around 3 points.

6.3.2 Importance of secondary language resources

To demonstrate the improvement in Hindi QG from the relatively larger English SQuAD dataset, we show results of using only HiQuAD during the main task in Table 5; unsupervised and supervised pretraining are still employed. We obtain modest performance improvements on the standard evaluation metrics (except ROUGE-L) by using English SQuAD data in the main task. These improvements (albeit small) demonstrate that our proposed cross-lingual framework is a step in the right di-
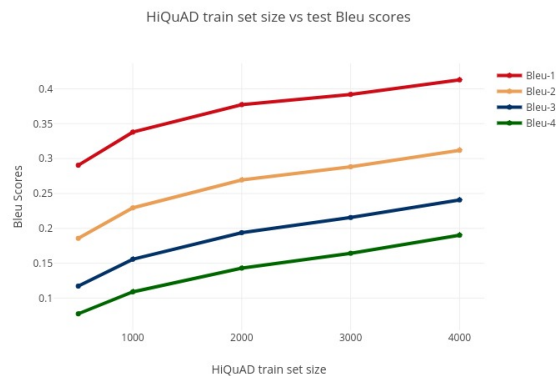

Figure 6: Trade-off between HiQuAD training dataset size and BLEU scores.

rection towards leveraging information from a secondary language.

6.4 How many sentence-question pairs are needed in the primary language?

To gain more insight into how much data is required to be able to generate questions of high quality, Fig. 6 presents a plot of BLEU scores when the number of Hindi sentence-question pairs is varied. Here, both unsupervised and supervised pretraining are employed but the English SQuAD dataset is not used. After significant jumps in BLEU-4 performance using the first 2000 sentences, we see a smaller but steady improvement in performance with the next set of 2000 sentences.

7 Conclusion

Neural models for automatic question generation using the standard sequence to sequence paradigm have been shown to perform reasonably well for languages such as English, which have a large number of training instances. However, large training sets are not available for most languages. To address this problem, we present a cross-lingual model that leverages a large QG dataset

in a secondary language (along with monolingual data and parallel data) to improve QG performance on a primary language with a limited number of QG training pairs. In future work, we will explore the use of cross-lingual embeddings to further improve performance on this task.

8 Acknowledgments

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *ICLR*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *ACL*.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Workshop on Machine Reading for Question Answering*.

Michael Heilman. 2011. Automatic factual question generation from text. *Language Technologies Institute School of Computer Science Carnegie Mellon University*.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *4th Workshop on Asian Translation (WAT2017)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Vishwajeet Kumar, Kireeti Boorla, Yogesh Meena, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018a. Automating reading comprehension by generating question and answer pairs. In *PAKDD*.

Vishwajeet Kumar, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018b. A framework for automatic question generation from text using deep reinforcement learning. *arXiv preprint arXiv:1808.04961*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The iit bombay english-hindi parallel corpus. In *LREC*.

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

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *NAACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL*.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*.

Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027*.

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 *NeurIPS*, pages 5998–6008.

Tong Wang, Xingdi Yuan, and Adam Trischler. 2017. A joint model for question answering and question generation. *arXiv preprint arXiv:1706.01450*.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *ACL*.

Hai-Tao Zheng, JX Han, JY Chen, and Arun Kumar Sangaiah. 2018. A novel framework for automatic chinese question generation based on multi-feature neural network model. *Comput. Sci. Inf. Syst.*