Does Meta-learning Help mBERT for Few-shot Question Generation in a Cross-lingual Transfer Setting for Indic Languages?

Aniruddha Roy[1], Rupak Kumar Thakur,[2*], Isha Sharma[1], Ashim Gupta[3],
Amrith Krishna[4], Sudeshna Sarkar[1] and Pawan Goyal[1]
[1]IIT Kharagpur, [2]Google India [3]University of Utah, [4]Uniphore
{aniruddha.roy,ishasharma}@iitkgp.ac.in,rupakthakur@google.com
sudeshna@cse.iitkgp.ac.in,pawang@cse.iitkgp.ac.in

Abstract

Few-shot Question Generation (QG) is an important and challenging problem in the Natural Language Generation (NLG) domain. Multilingual BERT (mBERT) has been successfully used in various Natural Language Understanding (NLU) applications. However, the question of how to utilize mBERT for few-shot QG, possibly with cross-lingual transfer, remains. In this paper, we try to explore how mBERT performs in few-shot QG (cross-lingual transfer) and also whether applying meta-learning on mBERT further improves the results. In our setting, we consider mBERT as the base model and fine-tune it using a seq-to-seq language modeling framework in a cross-lingual setting. Further, we apply the model agnostic meta-learning approach to our base model. We evaluate our model for two low-resource Indian languages, Bengali and Telugu, using the TyDi QA dataset. The proposed approach consistently improves the performance of the base model in few-shot settings and even works better than some heavily parameterized models in some settings. Human evaluation also confirms the effectiveness of our approach.

1 Introduction

QG can be defined as the task of generating an appropriate question based on the answer tokens and the context. The previous state-of-the-art QG models are built using neural networks (Du et al., 2017; Zhou et al., 2017; Zhao et al., 2018; Nema et al., 2019), and are trained on high-resource languages with availability of vast amount of manually annotated data for training. Collecting and annotating such vast data for training on low-resource languages can be challenging and costly. Cross-lingual transfer learning has shown its effectiveness in many NLP applications (Kumar et al., 2019; Chi et al., 2019; Asai et al., 2021; Xie et al., 2018) for addressing data scarcity, because it allows us to transmit domain knowledge from a high resource annotated source language to domain of desired target language by fine-tuning with data from a target domain with low resource availability. mBERT (Devlin et al., 2018) has been successfully used in various NLU tasks (Wu and Dredze, 2019; Hu et al., 2020). However, utilizing mBERT for generation tasks with cross-lingual transfer remains unexplored, specifically for QG.

In this paper, we examine the application of mBERT for QG with cross-lingual transfer. Specifically, we ask: 1) Despite the successful usage of various multilingual auto-regressive language models (Xue et al., 2020; Liu et al., 2020; Maurya et al., 2021), can mBERT, an encoder-based model with fewer parameters than these auto-regressive models, be used for QG with cross-lingual transfer? 2) In few-shot cross-lingual transfer settings, fine-tuning may cause colossal distribution gap and severe forgetting (French, 1999), along with an overfitting problem. Can applying meta-learning further improve the results? Meta-learning has shown its effectiveness in various NLP applications such as Dialogue Generation (Qian and Yu, 2019), Machine Translation (Park et al., 2021; Gu et al., 2018), and Natural Language Understanding (Nooralahzadeh et al., 2020; Roy et al., 2022) as it has the capacity to swiftly adapt to unseen training instances while leveraging limited resources, thus it may be helpful in this case as well.

To address these two questions, we use mBERT as the base model, and following (Dong et al., 2019), we fine-tune it as a sequence-to-sequence LM (unidirectional decoding conditioned on bidirectional encoding). We then apply the model agnostic meta-learning approach (Finn et al., 2017) to our base model, and we call our approach meta-QG. The goal of our proposed approach is to determine the best initialization of the model param-

*The author contributed to the paper when he was a student of IIT Kharagpur

Proceedings of the 29th International Conference on Computational Linguistics, pages 4251–4257
October 12–17, 2022.

eters for the QG task, which can help the model to easily adapt to target languages which are low-resource. In our method, there are two phases, i.e., meta-train phase and adaptation phase. The objective of the meta-train phase is to learn an optimal parameter initialization, so we create pseudo QG tasks on the source language. To minimize the language distribution gap between the meta-train and adaption phase, we mix English with an Indian language and consider both as the source languages. During the adaptation phase, we apply the model obtained using meta-train phase on the target language in zero-shot or few-shot settings. For evaluation, we apply our model on two low-resource Indian languages- Telugu and Bengali. We show that our approach gives consistent gains over the base model for Meteor, BLEU-4, and Rouge-L scores. Additionally, we also compare our approach with the heavily parameterized models mt5-base (Xue et al., 2020) (580M) and mBART-50 (Liu et al., 2020) (680M), and the results obtained demonstrate that our approach outperforms mt5-base for both the languages, and performs better than mBART-50 for Bengali in few-shot ($n \leq 16$) settings. Human evaluation also indicates that the proposed approach is very effective.

2 Methodology

QG task is defined as to generate a (syntactically and semantically correct) question based on a paragraph and the relevant sequence of answer tokens present in it. In our cross-lingual transfer setting, we denote the source language labelled training data as D^S_{train} and the target language test data as D^T_{test}. The aim of our QG meta-learning algorithm is to train a model with D^S_{train} using minimum or zero resource of target language labelled data, such that it performs well on D^T_{test}. Our base model in detail, as well as our proposed approach, is described below.

2.1 Base model

For the base model, we make use of multilingual BERT ($mBERT$) and fine-tune it (Dong et al., 2019) as a sequence-to-sequence LM for our QG task. In our work, we consider passage and answer as source segment and question as target segment, and we join these two segment with special tokens $[SEP]$. We randomly mask some tokens in the target sequence and fine-tune the model to recover the masked tokens in a sequence-to-sequence manner. Basically, the model considers partial sentence $y_1 : y_{t-1}$ from the ground truth (bidirectional encoding) to generate the t-th token y_t, which was masked (unidirectional decoding). We use beam search during decoding, taking beam size as 3.

2.2 Applying Meta-learning

Next, we discuss how we apply model-agnostic meta learning (Finn et al., 2017, MAML) for the proposed task. First we take the source languages and, using them, create a set of tasks which we refer to as pseudo-meta-QG tasks. Then, we train the base model on these using pseudo-meta-QG training. Lastly, we adapt the meta-trained model to the test examples of the target language (in zero-shot and few-shot settings). We discuss this in detail below.

Pseudo-meta-QG Tasks creation: We create pseudo-meta-QG tasks (Wu et al., 2020) from the source languages' labeled data. Let us assume that source language's training data, D^S_{train} has P examples denoted as $\{x^{(i)}\}^P_{i=1}$. For each example $x^{(i)}$, a pseudo-meta-QG task τ_i is created in the form of a pseudo train set $D^{\tau_i}_{train}$ and a test set $D^{\tau_i}_{test}$. Here, $D^{\tau_i}_{test} = x^{(i)}$, and $D^{\tau_i}_{train}$ is obtained by retrieving k examples from D^S_{train} which most closely resemble the selected test example. We use the input representation from the base model (mBERT) to calculate (cosine) similarity between any two examples. The pseudo-meta-QG tasks τ_i are defined as follows per training example:

$$\tau_i = (D^{\tau_i}_{train}, D^{\tau_i}_{test}), i \in 1, 2, ..., P. \quad (1)$$

Pseudo-meta-QG training setup: Given the base model M_θ (mBERT) with parameters θ and pseudo-meta-QG tasks $\{\tau_i\}^P_{i=1}$, we obtain θ'_i (one set of parameters per pseudo-meta-QG task τ_i) by doing an inner-update on each τ_i. Specifically, it performs few ($n = 2$) gradient steps on $D^{\tau_i}_{train}$ (pseudo train set), and helps to obtain new model parameters from the base model parameters θ. Our equation for inner-update is as follows:

$$\theta'_i = \theta - lr_{inner} \nabla_\theta \mathcal{L}^{\tau_i}_{D_{\text{train}}}(\theta) \quad (2)$$

Here, θ denotes parameters of the base model M_θ, lr_{inner} is inner learning rate, and $\mathcal{L}^{\tau_i}_{D_{\text{train}}}$ is the loss of pseudo training set $D^{\tau_i}_{train}$ of task τ_i. After the inner-update, a meta-update is performed on the

pseudo test set $D_{test}^{\tau_i}$ of τ_i . This step first calculates the pseudo test loss $\mathcal{L}_{D_{test}^{\tau_i}}$ by evaluation of the modified parameters (θ_i') on $D_{test}^{\tau_i}$. After that, we update the model by optimization of the loss on $D_{test}^{\tau_i}$ in terms of θ. There are multiple iterations involved in this step and the meta-update equation is defined as:

$$\begin{aligned} \theta &\leftarrow \theta - lr_{meta} \sum_i \nabla_\theta \mathcal{L}_{D_{\text{test}}}^{\tau_i}(\theta_i') \\ &= \theta - lr_{meta} \sum_i grad_i \end{aligned} \tag{3}$$

Here lr_{meta} is the learning rate of meta-update and $grad_i$ is the meta-gradient on task τ_i. We can expand it as:

$$grad_i = \nabla_\theta \mathcal{L}_{D_{\text{test}}}^{\tau_i}\left(\theta_i'\right) = \nabla_{\theta_i'} \mathcal{L}_{D_{\text{test}}^{\tau_i}}\left(\theta_i'\right) \nabla_\theta\left(\theta_i'\right) \tag{4}$$

In Equation 4, $\nabla_\theta(\theta_i')$ refers to the Jacobian matrix and it will introduce higher order gradient. Following (Finn et al., 2017; Wu et al., 2020), to reduce the computational cost, we use identity matrix in place of Jacobian matrix. Therefore, $grad_i$ can be computed as:

$$grad_i = \nabla_{\theta_i'} \mathcal{L}_{D_{\text{test}}^{\tau_i}}\left(\theta_i'\right) \tag{5}$$

Finally, we obtain the base model's updated parameters as θ^*.

Adaptation: In the adaptation phase, we apply the source trained model (parameters θ^*) to the target language's test samples in a zero-shot or few-shot setting. We follow Wu et al. (2020)'s adaptation approach for our zero-shot setting. In few-shot setting, we fine-tune the source-trained model on few-shot examples from the training data of target language. Specifically, we subsample the target language training dataset to obtain the small few-shot datasets of size [2,4,8,16]. We randomly sample five datasets for each shot.

3 Experiments

We evaluate our meta-learning based QG model in zero-shot and few-shot settings. This section covers details about the dataset used in our experiments followed by the implementation details with evaluation metrics. **Dataset:** We conduct experiments on low resource Indian languages having minimum amount of annotated data for QG. We use TyDi QA[1] (Clark et al., 2020) Gold passage dataset for our experiments. The dataset contains triplets of passage, question and answer for 9 languages. We evaluate our method on Bengali and Telugu dataset. The sizes of the Bengali and Telugu dataset (train, dev), in terms of number of examples, are (2390, 113), and (5563, 669), respectively. For cross-lingual knowledge transfer, we additionally use English triplets from the same dataset (train = 3696; dev = 440). One should note that since the aforementioned dataset contains no test data, we consider development set as test data for all our experiments. For evaluating Bengali, we consider English and Telugu as the source languages, while we use English and Bengali as the source languages for Telugu. The purpose of mixing one Indian languages is to learn different language distributions rather than single-source distribution. Please note that we follow the same zero-shot and few-shot approach to our base models for fair comparison.

Experiment Setup: We implement our algorithm using PyTorch 1.1.0. Our base model uses BERT base multilingual cased with 12 Transformer blocks, 12 self-attention heads and 768 hidden dimension, GELU activation, and dropout is 0.1. The maximum sequence length is set to 512 for the input. For the creation of pseudo-meta-QG task, we take only two $k = 2$ similar examples during meta training and zero-shot adaptation phase. Each meta-training step performs two inner-update and a meta-update on a batch of 16 tasks. We train our model up to 6,000 meta-training steps. As described in Wu and Dredze (2019), we freeze the embedding and the first three layers of the base transformer model, while the other layers are further fine-tuned for each task. Other hyper-parameter settings are same as in Devlin et al. (2018). We use Adam (Kingma and Ba, 2015) optimizer with learning rates of lr_{inner}, lr_{meta} = $3e{-}5$ for both inner-update and meta-update steps. We set learning rate of lr_{adapt} = $1e{-}5$ for gradient updates during adaptation phase. For few-shot experiment, we fine-tune the meta-trained model up to 60 steps.

We evaluate the systems using BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), and ROUGE-L (Lin, 2004) scores [2]. During the training phase, we train our model using the source language's training data and save the model based on the accuracy of the source lan-

[1] https://github.com/google-research-datasets/tydiqa

[2] We use (Du et al., 2017)'s script for evaluating our model

Model	Setting	Bengali			Telugu		
		BLEU-4	Meteor	Rouge-L	BLEU-4	Meteor	Rouge-L
mt5-base	0-shot	1.38	9.62	7.15	0.00	15.80	11.21
mBART-50		**4.31**	**20.92**	15.87	**3.52**	**27.15**	17.56
mBERT		3.24	16.37	27.88	2.27	17.82	15.03
meta-QG (Ours)		3.99	18.35	**29.45**	1.92	20.19	**20.19**
mt5-base	2-shot	1.73	13.80	9.97	1.15	21.96	12.56
mBART-50		5.01	**27.98**	21.00	**10.02**	**33.52**	**39.21**
mBERT		3.24	16.37	27.88	2.27	17.82	15.03
meta-QG (Ours)		**5.22**	25.45	**33.51**	4.86	31.77	31.83
mt5-base	4-shot	1.71	15.31	10.80	1.59	25.65	14.11
mBART-50		4.71	23.84	19.14	**10.38**	**36.04**	**37.12**
mBERT		3.24	16.37	27.88	2.27	17.82	15.03
meta-QG (Ours)		**5.38**	**26.23**	**34.48**	5.19	34.13	28.54
mt5-base	8-shot	2.95	19.52	13.81	3.88	29.25	19.25
mBART-50		5.01	**27.73**	20.91	**21.02**	**38.88**	**43.74**
mBERT		4.58	22.47	30.74	10.49	32.31	33.05
meta-QG (Ours)		**5.54**	27.40	**32.80**	10.19	36.58	34.57
mt5-base	16-shot	4.85	24.84	17.56	6.23	33.15	26.22
mBART-50		5.67	**27.91**	22.54	**26.46**	**39.44**	**50.72**
mBERT		6.35	23.39	33.99	12.05	32.75	34.94
meta-QG (Ours)		**8.45**	26.77	**37.17**	12.83	35.93	37.78

Table 1: Performance for zero-shot and few-shot cross-lingual question generation for Bengali and Telugu. We consider English and Telugu as source language and evaluate Bengali as target language, while for evaluation of target language Telugu we use English and Bengali as source language. The improvements in BLEU-4 by meta-QG were statistically significant ($p < 0.05$ as per $t-$test) for all settings wrt mt5-base and mBERT and for Bengali 16-shot setting wrt mBART-50.

guage's dev dataset. We carry out the training procedure for four random seeds. For few-shot setting, we randomly sample 5 datasets, and average over 4 training random seeds.

Results: In Table 1, we compare our model to the various base models in zero-shot and few-shot settings to verify the effectiveness of cross-lingual knowledge transfer from source languages to target languages. We see that meta-QG outperforms the base mBERT model for all the settings except Telugu 0-shot BLEU-4 and Telugu 8-shot BLEU-4. Interestingly, it also outperforms the heavily parameterized mt5-base model for all the settings. mBART-50, however, shows its superior quality and outperforms all the other methods for Telugu, except zero-shot Rouge-L, where meta-QG gives better scores. For Bengali, meta-QG still holds an edge over mBART-50, which was quite encouraging. The improvements in BLEU-4 by meta-QG were statistically significant ($p < 0.05$ as per $t-$test) for all settings wrt mt5-base and mBERT and for Bengali 16-shot setting wrt mBART-50 . A detailed error analysis is presented in the Appendix.

Human Evaluation: We also perform human evaluation using a similar procedure as used by (Chi et al., 2019; Maurya et al., 2021). We randomly sample 35 test data-points in both Telugu and Bengali languages and employ three metrics: fluency, relatedness, and correctness. Fluency measure is self-explanatory. The degree to which the generated questions are related to the input context is measured by relatedness, correctness assesses the meaning and semantics of the generated output. While fluency and correctness mainly deal with the generation quality, relatedness is the most critical among these for the task. We present the generated questions by all the competing models (after random shuffling) to three language experts and ask them to rate the questions on a 5-point Likert scale (1: very bad and 5: very good) for all the metrics. The results show that our approach consistently outperforms mBERT and mt5-base for all the metrics. mBART-50 achieves better scores in

Fluency and Correctness due to its superior generation capability. However, meta-QG performs better in relatedness for Bengali, the most critical metric. The final numbers are in Table 2 in the Appendix. These were calculated by averaging all the experts' responses for each parameter.

4 Conclusion

In this work, we make use of mBERT for QG task in few-shot cross-lingual transfer setting, and interestingly, we find that it actually performs better than mt5-base for all the settings, and better than mBART-50 for 16-shot setting in Bengali. We then explore the use of meta-learning with mBERT as the base model (meta-QG) and find that it achieves significant performance improvements compared to the mBERT as well as mt5-base, and surprisingly also outperforms mBART-50 for Bengali. In the future, we plan to extend this framework to other Natural Language Generation tasks, and also plan to study the effectiveness of data augmentation approaches.

Acknowledgement

We thank NLTM, BHASHINI, under the Ministry of Electronics and Information Technology, Govt of India, for their funding and support. We want to thank the human annotators who volunteered to be part of the human evaluation. We also thank Sovan Kumar Sahoo, Souvic Chakraborty, Anurag Roy, Santanu Pal, Ankit Bagde, Kousshik Raj, Nithish Kannen, and Nikhil Reddy, who helped in various aspects to complete the project.

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

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

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

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

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*.

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.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4*, volume 1:, page 13421352.

C. ; Abbeel Finn, P. ;, and S. Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML, 11261135*.

Robert French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3:128–135.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor O. K. Li. 2018. Meta-learning for low-resource neural machine translation.

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

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

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, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. Zmbart: An unsupervised cross-lingual transfer framework for language generation.

Preksha Nema, Akash Kumar Mohankumar, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2019. Let's ask again: Refine network for automatic question generation.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Cheonbok Park, Yunwon Tae, Taehee Kim, Soyoung Yang, Mohammad Azam Khan, Eunjeong Park, and Jaegul Choo. 2021. Unsupervised neural machine translation for low-resource domains via meta-learning.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649, Florence, Italy. Association for Computational Linguistics.

Aniruddha Roy, Isha Sharma, Sudeshna Sarkar, and Pawan Goyal. 2022. Meta-ed: Cross-lingual event detection using meta-learning for indian languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources.

S. Wu and M. Dredze. 2019. *Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT*. CoRRabs/1904.09077.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910, Brussels, Belgium. Association for Computational Linguistics.

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

A Appendix

A.1 Human Evaluation Results: Complete Table

Model	Flu.		Rel.		Cor.	
	bn	te	bn	te	bn	te
mt5-base	3.17	2.76	2.35	2.80	3.28	2.65
mBART-50	**4.42**	**4.32**	2.69	**3.23**	**4.29**	**3.91**
mBERT	3.01	3.40	2.17	2.49	2.74	3.08
meta-QG	3.49	3.49	**2.96**	2.85	3.79	3.51

Table 2: Human evaluation results of 16-shot cross-lingual question generation for Bengali and Telugu. The three metrics are Fluency (Flu.), Relatedness (Rel.), and Correctness (Cor.) respectively.

A.2 Case Study

Table 3 shows few example sentences with the corresponding questions generated by the base mBERT model as well as the proposed meta-QG approach. For the examples 3a and 3d, we find that mBERT does not generate a question where entity names are getting repeated, possibly due to some bias towards generating entity names from the reference context. However, meta-QG overcomes this issue and generates better questions. The questions generated by mBERT in 3b and 3c are better than the other two examples, but there are minor issues, such as 'Surya Sen' instead of 'Surya Sen's' (3b: missing morphological marker in Bengali) and only the surname (3c).

Reference 3a (Bengali): চিত্রা বন্দ্যোপাধ্যায়ের স্বামীর নাম কী ? **Translation**: What is the name of Chitra Bandyopadhyay's husband? **meta-QG output**: চিত্রা বন্দ্যোপাধ্যায়ের স্বামীর নাম কী ? **Translation**: What is the name of Chitra Bandyopadhyay's husband? **mBERT output**: চিত্রা বা চিত্রা বা চিত্রা ছিলেন ? **Translation**: Was it Chitra or Chitra or Chitra?
Reference 3b (Bengali): মাস্টারদা সূর্যকুমার সেনের বাবার নাম কী ছিল ? **Translation**: What was the name of Masterda Suryakumar Sen's father? **meta-QG output**: সূর্য সেনের বাবার নাম কী ? **Translation**: What is the name of Surya Sen's father? **mBERT output**: সূর্য সেন বাবার নাম কী ? **Translation**: What is the name of Surya Sen father?
Reference 3c (Bengali): বিখ্যাত জ্যোতির্বিজ্ঞানী নিকোলাউস কোপের্নিকুসের জন্ম কবে হয় ? **Translation**:When was the famous astronomer Nicolaus Copernicus born? **meta-QG output**: নিকোলাস কোপারনিকাস জন্ম কবে ? **Translation:** When was Nicholas Copernicus born? **mBERT output** : কোপারনিকাসের জন্ম কবে ? **Translation:** When was Copernicus born?
Reference 3d (Bengali): বিখ্যাত বাংলাদেশী চলচ্চিত্র পরিচালক মোরশেদুল ইসলামের প্রথম পরিচালিত চলচ্চিত্রের নাম কী ? **Translation:** What is the name of the first film directed by famous Bangladeshi film director Morshedul Islam? **meta-QG output**: মোরশেদুল ইসলামের প্রথম চলচ্চিত্রের নাম কী ? **Translation:** What is the name of the first film of Morshedul Islam? **mBERT output** : মোরশেদুল ইসলাম বা মোরশেদুল ইসলাম বা মোরশেদ কে ছিলেন ? **Translation:** Who was Morshedul Islam or Morshedul Islam or Morshed?

Table 3: Some example outputs by the base mBERT model as well as the proposed meta-QG approach