EXPLOITING ADAPTERS FOR QUESTION GENERATION FROM TAMIL TEXT IN A ZERO-RESOURCE SETTING

Shanmukanathan Purusanth

(209367E)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

Faculty of Engineering

University of Moratuwa

Sri Lanka

October 2022

EXPLOITING ADAPTERS FOR QUESTION GENERATION FROM TAMIL TEXT IN A ZERO-RESOURCE SETTING

Shanmukanathan Purusanth

(209367E)

This dissertation submitted in partial fulfillment of the requirements for the Degree of MSc in Computer Science specializing in Data Science

Department of Computer Science and Engineering

Faculty of Engineering

University of Moratuwa

Sri Lanka

October 2022

DECLARATION

I declare that this is my own work, and this dissertation does not incorporate without acknowledgment, any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to the University of Moratuwa, the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic, or another medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for the evaluation of a Master's thesis.

Name of the Supervisor: Dr. Surangika Ranathunga

Signature of the supervisor: Date

ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to Dr. Surangika Ranathunga for guiding me to initiate, finding a great research topic to conduct the research, and for continuing support and encouragement. Her supervision greatly helped me in setting goals and engaging in the research.

I would like to express my greatest gratitude to the Department of Computer Science and Engineering, the University of Moratuwa for providing the support to overcome this effort. Last but not least, my heartfelt gratitude goes to my parents and friends who supported me throughout this effort.

ABSTRACT

Automatic Question Generation focuses on generating questions from a span of text is a significant problem in Natural Language Processing (NLP). Question generation in low-resource languages is under-explored compared to high-resource languages. In the earlier work, all the parameters of a pre-trained multilingual language model were fine-tuned to perform a zero-shot question generation and other sequence-to-sequence (S2S) generation tasks. However, such full model fine-tuning is not computationally efficient. Recent research introduced a neural module called adapter to each Transformer layer of a pre-trained language model and fine-tuned only the adapter parameters to mitigate this issue. In this study, we explored single task adapter and adapter fusion on the pre-trained multilingual model mBART to generate questions from Tamil text. Our best model produced a Rough-1 (F1) score of 16.9. Furthermore, we obtained a similar result with two variants of adapters called Houlsby adapter [1] and Pfeifer adapter [1], which resemble the result of adapters for other S2S tasks[2].

Keywords: Automatic Question Generation, Pre-trained language models, Adapters, Tamil.

TABLE OF CONTENTS

Declaration of the candidate and the supervisor I
Acknowledgements II
Abstract III
Table of content IV
List of Figures VI
List of Tables VII
List of Abbreviation VIII
1. Introduction 1
1.1 Background 1
1.2 Research Problem 1
1.3 Research Questions 2
1.4 Objective 2
1.5 Thesis Structure 3
2. Literature Review 4
2.1 Overview 4
2.2 Datasets 4
2.3 Rule-based approach 4
2.4 Machine Learning-based approach 5
2.4.1 Single-hop question generation 6
2.4.2 Multi-hop question generation 8
2.5 Pre-trained sequence-to-sequence modes 8
2.5.1 Adapters and Adapter Fusion 9
2.6 Summary 10
3. Methodology 11
3.1 Experimental setups 11
3.2 Baseline 12
3.2.1 Full model fine-tuning(mBART50) 12
3.2.2 Language model fine-tuning(T5) 12
3.2.3 Freeze token embedding and position of mBART50 12
3.2.4 Zero-shot questions generator 12
3.2.5 Multiple adapters for cross-lingual transfer learning 12
3.2.6 Prefix tuning 13
3.3.1 Proposed Method 13
3.3.1 Adapter fine-tuning with Houlsby adapter 13

3.3.2 Adapter fine-tuning with Pfeiffer adapter 13
3.3.3 Adapter Fusion 13
4. Evaluation Results 14
4.1 Evaluation metrics 14
5.1 Datasets 14
5.2 Results 15
6. Conclusion and Future work 18
Reference List 19

LIST OF FIGURES

Figure	description	Page
Figure 2.4.1	BERT-QG architecture	7
Figure 2.5.1	Houlsby adapter	9
Figure 2.5.2	Pfeifer adapter	9
Figure 2.5.3	Architecture of Adapter fusion inside a Transformer layer	11

LIST OF TABLES

Table	Description	Page
Table 1.1	Example of a question generated by MulQG	1
Table 2.4.1	Distribution of question words in SQuRD and MARCO dataset	7
Table 3.1.1	An example of formatted input and its formatted target for mBART	11
Table 3.1.2	An example of formatted input and its formatted target for mT5	11
Table 4.2.1	An example question form the test dataset	15
Table 4.2.2	Average lengths of context, question, answer	15
Table 4.3.1	Models and their parameters and trained epochs	16
Table 4.3.2	Result	17

LIST OF ABBREVIATION

Abbreviation	Description
S2S	sequence-to-sequence
SOTA	state-of-the-art
POST	part-of-speech tagging
NLP	Natural language processing