Pre-training and Fine-tuning Multilingual Sequence-To-Sequence Models for Domain-Specific Low-Resource Neural Machine Translation

Sarubi Thillainathan

208037K

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

April 2022

DECLARATION

I, Sarubi Thillainathan, declare that this is my own work and this dissertation does not incorporate without acknowledgement 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 acknowledgement is made in the text.

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

Signature: Date:

The above candidate has carried out research for the Masters thesis/Dissertation under my supervision.

Name of Supervisor: Dr. Surangika Ranathunga

Signature of the Supervisor: Date:

Name of Supervisor: Prof. Sanath Jayasena

Signature of the Supervisor: Date:

ABSTRACT

Limited parallel data is a major bottleneck for morphologically rich Low-Resource Languages (LRLs), resulting in Neural Machine Translation (NMT) systems of poor quality. Language representation learning in a self-supervised sequence-to-sequence fashion has become a new paradigm that utilizes the largely available monolingual data and alleviates the parallel data scarcity issue in NMT. The language pairs supported by the Self-supervised Multilingual Sequence-to-sequence Pre-trained (SMSP) model can be fine-tuned using this pre-trained model with a small amount of parallel data.

This study shows the viability of fine-tuning such SMSP models for an extremely low-resource domain-specific NMT setting. We choose one such pre-trained model: mBART. We are the first to implement and demonstrate the viability of non-English centric complete fine-tuning on SMSP models. To demonstrate, we select Sinhala, Tamil and English languages in extremely low-resource settings in the domain of official government documents.

This research explores the ways to extend SMSP models to adapt to new domains and improve the fine-tuning process of SMSP models to obtain a high-quality translation in an extremely low-resource setting. We propose two novel approaches: (1) Continual Pre-training of the SMSP model in a self-supervised manner with domain-specific monolingual data to incorporate new domains and (2) multistage fine-tuning of the SMSP model with in- and out-domain parallel data.

Our experiments with Sinhala (Si), Tamil (Ta) and English (En) show that directly fine-tuning (single-step) the SMSP model mBART for LRLs significantly outperforms state-of-the-art Transformer based NMT models in all language pairs in all six bilingual directions. We gain a +7.17 BLEU score on Si→En translation and a +6.74 BLEU score for the Ta→En direction. Most importantly, for non-English centric Si-Ta fine-tuning, we surpassed the state-of-the-art Transformer based NMT model by gaining a +4.11 BLEU score on Ta→Si and a +2.78 BLEU score on Si→Ta.

Moreover, our proposed approaches improved performance strongly by around a +1 BLEU score compared to the strong single-step direct mBART fine-tuning for all six directions. At last, we propose a multi-model ensemble that improved the performance in all the cases where we obtained the overall best model with a +2 BLEU score improvement.

Keywords: Neural Machine Translation, Pre-trained Language Models, Pre-training, Fine-tuning, Low-Resource languages, mBART

DEDICATION

With deepest gratitude, I dedicate this research to my Grandpa, *Late M. Saravanamuttu*. And, of course, to my forever loving family for supporting my dreams, no matter what!

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisors Dr Surangika Ranathunga and Prof. Sanath Jayasena, for the continuous support and guidance given for the success of this research. Without your insights and tremendous mentorship, I could not have achieved this level. Thank you for always stepping in to help whenever I needed regardless of your busy schedules.

I wish to convey my sincere appreciation to Prof. Gihan Dias and National Language Processing Center (NLPC) members for their valuable insights and support given for this research. I would also like to thank the entire Department of Computer Science and Engineering staff, both academic and non-academic, for their help and for providing me with the resources necessary to conduct my research. This research was supported by the University of Moratuwa AHEAD project Research Grant.

Lastly, I want to thank my family and friends who supported me though this journey.

Thank you!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
MT	Machine Translation
SMT	Statistical Machine Translation
NMT	Neural Machine Translation
MNMT	Multilingual Neural Machine Translation
RNN	Recurrent Neural Networks
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
DAE	Denoising Autoencoder
MLE	Maximum Likelihood Estimate
TL	Transfer Learning
SMSP	Self-supervised Multilingual Sequence-to-sequence Pre-trained
FT	Fine-Tuning
LRL	Low-Resource Language
LM	Language Model
M-FT	Multilingual Fine-Tuning
B-FT	Bilingual Fine-Tuning
CPT	Continual Pre-Training