Translation of Named Entities Between Sinhala and Tamil for Official Government Documents

Thayaparan Mokanarangan

(178089G)

Degree of Master of Science (Research)

Department of Computer Science And Engineering

University of Moratuwa
Sri Lanka

August 2018

Translation of Named Entities Between Sinhala and Tamil for Official Government Documents

Thayaparan Mokanarangan

(178089G)

Thesis submitted in partial fulfillment of the requirements for the
Degree of Master of Science (Research) in Computer Science and Engineering

Department of Computer Science And Engineering

University of Moratuwa
Sri Lanka

August 2018

Declaration

I, Thayaparan Mokanarangan, 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).

Signed: _____

Date: _____

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

Name of Supervisor: Dr. Surangika Ranathunga

Signature of supervisor: _____ Date: _____

Name of Supervisor: Dr. Uthayasanker Thayasivam

Signature of supervisor: _____ Date: _____

"It was awesome, but also.. it wasn't?"

Troy from Community

UNIVERSITY OF MORATUWA

Abstract

Faculty Of Engineering
Department of Computer Science And Engineering

Master of Science

by Thayaparan Mokanarangan
(178089G)

Analyzing existing machine translation approaches for Sinhala-Tamil official government documents have revealed the shortcomings when translating named entities. The diverse nature of the domain coupled with the lack of resources and morphological complexity are the key reasons for this problem. Our research focuses on translating named entities for official government documents between Tamil and Sinhala. In this research, we focus on identifying and translating named entities to improve the translation performance. We present a novel tag set specific to official government documents and also propose a graph-based semi-supervised approach that works better than state-of-the-art approaches for low-resource settings. We employed this approach to build a large annotated corpus in a cost-effective manner from a smaller amount of seed data and was able to build an annotated corpus of over 200K words each for Tamil and Sinhala. We also implemented a deep-learning approach for Named Entity Recognizer that gave the best output for a completed corpus. Since the deep-learning approach was a generic solution for sequential tagging, we also employed it to build a Part-of-Speech tagger that outperforms existing systems. The University of Moratuwa already has a system for translating official government documents called *SiTa*. Finally, we incorporated the aforementioned models to build a module that translated named entities and integrated it to *SiTa*. We empirically show that our modules improve over the baseline for Tamil $\rightarrow$ Sinhala and Sinhala $\rightarrow$ Tamil translation tasks by upto 0.5 and 1.4 BLEU scores, respectively.

Keywords: Machine Translation. Named Entity Recognition, Graph-Based Semi-Supervised Learning, Deep Learning, Named Entity Translation

Acknowledgements

I would never have been able to finish my dissertation without the guidance, support and encouragement of numerous people including my mentors, my friends, colleagues and support from my family. At the end of my thesis I would like to thank all those people who made this thesis possible and an unforgettable experience for me.

First and foremost, I would like to express my sincere gratitude to my supervisors Dr. Surangika Ranthunga and Dr. Uthayasanker Thayasivam, for the continuous support given for the success of this research both in unseen and unconcealed ways. This would not have been a success without your tremendous mentorship and advice from the beginning. Your wide knowledge and logical way of thinking have been of great source of inspiration for me. You have always extended his helping hands in solving research problems. The in-depth discussions, scholarly supervision and constructive suggestions received from you have broadened my knowledge. I strongly believe that without your guidance, the present work could have not reached this stage.

I wish to thank Prof. Gihan Dias and Prof. Sanath Jayasena for their supervision, advice, and guidance from the very early stage of this research as well as giving me extraordinary experiences through-out the work. This research was supported by the Department of Official Languages and the University of Moratuwa Senate Research Grant. I sincerely thank the colleagues from the Department of Official Languages for the support given.

I would like to thank Ms. Fathima Farhath, Ms. Nimasha Dilshani and Ms. Yashothara Shanmugarajah, who as good friends from my graduate studies, were always willing to help and give their best suggestions.

Thank you!

Contents

List of Figures

List of Tables

Abbreviations

NLP	**N**atural **L**angaguge **P**rocessing
NER	**N**amed **E**ntity **R**ecognition
NE	**N**amed **E**ntities
BLEU	**B**i-**L**ingual **E**valuation **U**nderstudy
MT	**M**achine **T**ranslation
CRF	**C**onditional **R**andom **F**ield
LSTM	**L**ong **S**hort **T**erm **M**emory
GSSL	**G**raph **B**ased **S**emi-**S**upervised **L**earning
CBOW	**C**ontinous **B**ag **O**f **W**ords
ME	**M**aximum **E**ntropy
HMM	**H**idden **M**arkov **M**odel
POS	**P**arts **O**f **S**peech
SMT	**S**tatistical **M**achine **T**ranslation
NMT	**N**eural **M**achine **T**ranslation