EFFICIENT TRANSLATION BETWEEN SINHALA AND TAMIL Using Part of Speech and Morphology

Yashothara Shanmugarasa

(178029B)

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 & Engineering

University of Moratuwa

Sri Lanka

November 2018

DECLARATION

“I declare that this dissertation has been composed by solely by myself and this dissertation does not combine without acknowledgement of 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:

Name: Yashothara.S

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

Signature of the supervisor: Date:

Name of the supervisor: Prof. Gihan Dias

Signature of the co-supervisor: Date:

Name of the supervisor: Dr. R.T. Uthayasanker

Abstract

Machine translation is the process of translating a document from one language to another with the aid of a computer. Even though many machine translation technologies exist, statistical machine translation (SMT) still provides better performance in terms of quality and time for low resourced languages. In this study, we choose Sinhala to Tamil translation and vice versa since they are official languages of Sri Lanka. Often government official documents are written in one language (the majority in Sinhala) and translated into Tamil. Translation between Tamil and Sinhala is currently a time-consuming manual process carried out by department of official languages. Aiding the translators with an automated translation system will improve the efficiency of this process. However, there are some challenges in statistical machine translation of Tamil and Sinhala.

In the initial part of this research, a study on language divergence was conducted to identify the challenges in machine translation between Tamil and Sinhala. In this thesis, we focus on (i) improving the statistical machine translation from Sinhala to Tamil using a hierarchical phrase-based SMT model, (ii) Parts of Speech (POS) based Factored Statistical Machine Translation system (F-SMT) and (iii) preprocessing techniques based on chunking and segmentation. Based on the analyzed results of language divergence, translation challenges such as (i) reordering, (ii) abbreviations and initials, (iii) word flow of the sentence, (iv) data sparseness, (v) ambiguity in translation, (vi) divergence among Tamil and Sinhala POS tagsets and (vii) mapping one word with one or more words were addressed. We also developed an algorithm for the alignment of different POS tagsets.

Subsequently, we used hierarchical phrase-based model and Factored model with POS integration to address challenges such as word reordering, word flow, context aware word selecting, translating conjunction words, better word choice and translating initials and abbreviations. Further we experimented with some pre-processing techniques based on chunking and segmentation towards addressing challenges such as unknown words, context awareness, better word choice, word flow, ambiguity in translation, translating into proper ‘Sandhi’ form, translating named entities and replacing one word with multiple words. Point-wise Mutual Information (PMI) based collocation phrases, POS based chunks, Named Entities and sub word segments are used to enhance the preprocessing step. Even though, the standard structure of a sentence is Subject-Object-Verb in both languages, there is a need of reordering in the translation between these languages. As our languages are the low resourced when we try to translate using traditional Statistical machine translation, we are unable to get a good order of sentences because of sub-phrases which have been observed previously in the training corpus can only reorder using distortion reordering model which is independent of their context. To improve reordering, we have tried the hierarchical phrase-based model and factored model. Hierarchical Phrase-based Model helps to improve translation quality between languages that vary by sentence structure. But it lowers the quality of languages share similar sentence structure and Tamil and Sinhala languages don’t have a syntactic parser for better performance.

Parts of speech knowledge is added as the factored model to improve reordering also. The words are factored into lemma and parts of speech. This factored model decreases the data sparseness in decoding and helps to reordering. These linguistic features are considered as separate tokens in the training process. We show that by generalizing translation with parts of speech tags, we could improve performance by 0.74 BLEU on a small Sinhala-Tamil system. Even though we could only achieve small increment in BLEU score, manual evaluation of the translation showed improvements.

Preprocessing is another way of enhancing the quality of the translation. Preprocessing described in the research is related to finding collocation words from PMI, NER based chunking, POS based chunking and segmentation. We observed that each of the preprocessing techniques provided better performance than the baseline system. When comparing the preprocessing methods, PMI based chunking gave good results compared to other preprocessing techniques. A hybrid approach is done by combining preprocessing approaches based on PMI chunking, NER chunking, and POS chunking. BLEU score was increased up to 33.41 by using a hybrid approach. The best performance is reported with hybrid approach for Sinhala to Tamil translation. We could improve performance by 12% BLEU (3.61) using a small Sinhala to Tamil corpus with the help of proposed hybrid approach preprocessing technique. Notably, this increase is significantly higher compared to the increase shown by prior approaches for the same language pair.

Keywords- Statistical Machine Translation, Parts of Speech, POS tagset Mapping, POS tagset Alignment, Semi-Supervised Approach, BIS tagset, UOM tagset, Tamil NLP, Sinhala NLP, Hierarchical Phrase Based model, Parallel corpus

ACKNOWLEDGEMENT

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 convey my sincere gratitude to my supervisors Dr. R.T. Uthayasanker and Prof. Gihan Dias, 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. Their wide knowledge and logical way of thinking have been of great source of inspiration for me. They have always extended their helping hands in solving research problems. I strongly trust that without their guidance, the present work could have not reached this stage.

I wish to thank Dr.Surangika Ranathunga 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 Department of Official Languages and University of Moratuwa Senate Research Grant. I sincerely thank colleagues from Department of Official Languages for the support given.

I would like to thank all staff from the Department of Computer Science and Engineering for their kindness expressed in all occasions.

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

Special thanks to my loving family. Your encouragement always motivated me to do my best. I would also like to thank all my friends who encouraged me all times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AU-KBC – Anna University K B Chandrasekhar

BIS – Bureau Indian Standard

BL – Base Line

BLEU – Bi-Lingual Evaluation Understudy

CIIL – Central Institute of Indian Languages

CRF– Conditional Random Fields

EBMT– Example based Machine Translation

EM – Expectation Maximization

F-SMT – Factored Statistical Machine Translation

HPB – Hierarchical Phrase Based

IIIT– International Institute of Information Technology

MT– Machine Translation

NIST – National Institute of Standards and Technology

NLP – Natural Language Processing

POS– Parts of Speech

RBMT– Rule based Machine Translation

SCFG – Synchronous Context-Free Grammar

SMT – Statistical Machine Translation

SOV– Subject-Object-Verb

SRILM – Stanford Research Institute for Language Modeling

SVM– Support Vector Machine

TER– Translation Edit Rate

TnT– Trigrams n Tagger

UCSC – University of Colombo School of Computing

UOM – University of Moratuwa

WER– Word Error Rate