

**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

DECLARATION	i
Abstract	ii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS	v
LIST OF TABLES	x
LIST OF FIGURES	xi
LIST OF ABBREVIATIONS	xiii
INTRODUCTION	1
1.1 Overview of machine translation	1
1.1.2 Importance and Application	3
1.2.4 Problem Definition	4
1.3 Motivation of the Thesis	5
1.3.1 Research motivation	5
1.3.2 Si-Ta system motivation	7
1.4 Objective of the Thesis	7
1.5 Contributions	7
1.5.1. Articles	8
1.5.2. Algorithms	9
1.6 Research Methodology	9
1.7 Organization of the Thesis	11
LITERATURE SURVEY	12
2.1. Overview	12
2.2 Various Approaches in Machine Translation	13

2.2.1 Rule based or Linguistic approach.....	14
2.2.2 Non-Linguistic Approaches	14
2.2.3 Hybrid machine translation system.....	17
2.3. Existing machine translation systems for Sinhala and Tamil languages	17
2.4 Existing Approaches in Language Divergence.....	23
2.5 Literature review about the available POS tagsets and linguistic tools for both Tamil and Sinhala languages	25
2.5.1 Tamil Language POS Tagsets.....	26
2.5.2 Tamil Language POS Tagger.....	28
2.5.3 Sinhala Language POS Tagsets	31
2.4.4 Sinhala Language POS Taggers.....	32
2.6 Existing Approach for Alignment between Different POS Tagsets	34
2.6.1 Existing Approaches on POS Standardization.....	34
2.6.2 Existing Approaches on Mapping From Different Tree-Bank Tagsets To Universal Set.....	36
2.7. Existing machine translation systems using hierarchical phrase-based model.....	37
2.8 Existing machine translation systems using factored phrase-based model	40
2.9 Existing Translation Systems Using Chunking the Words.....	42
2.10 Existing Translation Systems Using Segmenting the Words	43
THEORETICAL BACKGROUND.....	46
3.1. General.....	46
3.1.2 Morphological Richness of Sinhala/Tamil Language.....	46
3.1.3 Challenges in Tamil/Sinhala Translation.....	47
3.2 Language Divergence	50
3.2.1 Dorr’s classification	51

3.3 POS Alignment	51
3.4 Statistical Machine Translation.....	53
3.4.1 Formalism of Statistical Machine translation	56
3.4.2 Architecture of Statistical Machine Translation	56
Translation Model.....	57
Language Model	58
Here I have explained language model for 1-gram. But in this research I have used 3-gram model for the experiments I have done.	59
The Statistical Machine Translation Decoder	59
3.4.3 Common challenges of SMT system	60
3.4 Hierarchical Phrase Based Translation	63
3.5 Factored Model	64
3.6 Chunking.....	66
3.7 Segmentation.....	68
3.8 Evaluating Statistical Machine Translation	69
3.8.1 Human Evaluation Techniques	69
3.8.2 Automatic Evaluation Techniques	70
3.9 Si-Ta System.....	71
3.10 Summary	74
METHODOLOGY	75
4.1 Language divergence between Sinhala and Tamil languages.....	76
4.2 Semi-Automatic Alignment of Multilingual Parts of Speech Tagsets.....	76
4.2.1 Tagset Selection	77
4.2.2 Semi-automatic algorithm for POS Tagset Alignment.....	80
4.3 Hierarchical Phrase-based model Machine Translation.....	81

4.3.1 Baseline system.....	81
4.3.2 Hierarchical Model	82
4.4 POS Integration to SMT system	82
4.4.1 Automatic Creation of Factored Corpora	83
4.4.2 Factored SMT for Sinhala and Tamil Language.....	84
4.5 Preprocessing based on Chunking	86
4.5.1 PMI based preprocessing	87
4.5.2 NER based chunking preprocessing	88
4.5.2 POS based chunking preprocessing	89
4.6 Preprocessing based on Segmentation	90
EXPERIMENTS	92
5.1 Overview	92
5.2 Language Divergence	93
5.3 Semi-automatic algorithm for aligning different POS tagsets.....	95
5.4 Hierarchical phrase based machine translation	97
5.4.1 Dataset.....	97
5.4.2 Experimental setup.....	98
5.5 POS Integration to SMT system	99
5.5.1 Dataset.....	99
5.5.2 Experimental setup.....	100
5.6 Preprocessing based on chunking	103
5.6.1 Dataset.....	104
5.6.2 Experimental setup for PMI based chunking	104
5.6.3 Experimental setup for NER based chunking	107

5.6.4 Experimental setup for POS based chunking.....	108
5.7 Preprocessing based on segmentation.....	110
5.7.1 Dataset.....	110
5.7.2 Experimental setup for segmenting the words into sub-word.....	111
5.8 Tamil to Sinhala traditional SMT system	112
5.8.1 Dataset.....	112
5.8.2 Experimental setup for Tamil to Sinhala SMT	113
RESULTS AND DISCUSSION	115
6.1 Language Divergence	115
6.1.1 Conflational Divergence	115
6.1.2 Inflectional Divergence.....	116
6.1.3 Categorical Divergence.....	117
6.1.4 Lexical Divergence	118
6.2 Semi-automatic alignment between Tamil and Sinhala POS tagsets	125
6.2.1 Equal relationship	129
6.2.2 Subsumption relationship.....	129
6.2.3 Complex relationship	132
6.3 Hierarchical phrase-based model machine translation system	133
6.4 POS Integration to SMT system	136
6.4.1 Human Evaluation	138
6.5 Preprocessing based on chunking	144
6.5.1 Results for PMI based chunking	144
6.5.2 Results for NER based chunking and POS based chunking.....	148
6.6 Preprocessing based on segmentation.....	149

6.7 Tamil to Sinhala traditional SMT system	150
6.7 Summary	151
CONCLUSION AND FUTURE WORK	152
7.1 Summary	152
7.2 Conclusion	154
8.2 Future Directions	155
REFERENCES	157

LIST OF TABLES

Table 2.1 Details and comparison of existing Sinhala-Tamil translation systems in terms of corpus size, domain, accuracy and linguistic information.....	22
Table 3.1 Snippet of a phrase translation table	58
Table 3.2 Snippet of the Language model	59
Table 3.3.PMI score between two adjacent words	67
Table 3.4 Scales of Manual Evaluation	70
Table 4.1 UOM tagset in two levels	78
Table 4.2 BIS tagset in two levels.....	79
Table 4.3 Factored Parallel Sentences in Sinhala and Tamil.....	84
Table 4.4 Three kinds of translation model and LM	85
Table 5.1 Complete Statistics of Parallel Corpus (In Sentence).....	98
Table 5.2 Sources of parallel data.....	100
Table 5.3 Tamil Monolingual Data.....	100
Table 5.4 Three kinds of translation model and LM in POS integration.....	101
Table 5.5 Statistics of training, tuning, testing and language model	104
Table 5.6 Statistics of training, tuning, testing and language model	111

Table 5.7 Sources of parallel data	113
Table 5.8 Sinhala Monolingual Data	113
Table 6.1 The case markers and the postpositions.....	121
Table 6.2 Tenses and Examples	124
Table 6.3 Divergence of the determiner system of Sinhala and Tamil.....	124
Table 6.4. Alignment of BIS tagset and UOM tagset	126
Table 6.5 Comparison of BLEU evaluation score with traditional Phrase-based model	133
Table 6.6 Some examples of translation generated by the translation system developed in this study	135
Table 6.7 The results are in terms of BLEU score for Baseline and POS integrated models	137
Table 6.8 4 point scale system for human evaluation	138
Table 6.9 Results of the comparison between the translations which are different between POS integrated system and Baseline systems.....	140
Table 6.10 The results are in terms of BLEU score for all models.....	145
Table 6.11 Sample translations of both models	146
Table 6.12 The results are in terms of BLEU score for all models.....	148
Table 6.13 Sample translations of both models	148
Table 6.14 BLEU Score values of the traditional phrase-based and fully segmented approaches.....	149

LIST OF FIGURES

Figure 1.1 Overview of Statistical Machine Translation: Learning Patterns from the parallel corpus	2
Figure 2.1 Direct, Interlingua and transfer approaches in Machine Translation	14

Figure 2.2 Block diagram of SMT system. S=Source language sentences, T=Target language sentences, TM=Translation model, LM=Language model	16
Figure 3.1 Snippet of the alignment between Tamil and Sinhala languages	53
Figure 3.2 Alignment of phrases of both languages E and T.....	57
Figure 3.3 Decoding process of Statistical Machine Translation in terms of Sinhala to Tamil translation	60
Figure 3.4 Tamil to English translation showing reordering. S=Sentence	64
Figure 3.5 Redefining a word from a single symbol to a vector of factors	65
Figure 3.6 Blocked diagram of Factored translation	65
Figure 3.7 The standard workflow for Morfessor command line tools	69
Figure 3.8 Architecture of Si-Ta System	71
Figure 3.9 Workflow diagram of Si-Ta System.....	72
Figure 3.10 User interface of Si-Ta System.....	73
Figure 4.1 Work flow of the semi-automatic POS tagsets alignment of P1 and P2 languages. T1=POS tagged data in P1 language, T2= POS tagged data in P2 language	80
Figure 4.2 Mapping Sinhala factors to Tamil Factors	84
Figure 4.3 Workflow of POS integrated SMT system.....	86
Figure 4.4 Phrase based statistical machine translation system with Preprocessing	87
Figure 6.1 Working of Hierarchical Phrase based decoder for Tamil to English translation.....	135
Figure 6.2 Graph of various machine translation models and the BLEU score	137
Figure 6.3 Pie chart for the sentences which are same and different between POS integrated model and Baseline	139
Figure 6.4 Pie chart for Comparative results related to human reference in the same category.....	140
Figure 6.5 Pie chart for Comparative results of better translations belong to POS	141

Figure 6.6 Pie chart for Comparative results of better translations belong to baseline model related to human reference	141
Figure 6.7 Graph of various different sizes of chunked words based on PMI score and the BLEU score	145

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
IIT– 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