Sinhala - Tamil Statistical Machine Translation (SMT) for Official Documents

Farook Fathima Farhath

(168034C)

Degree of Master of Philosophy in Computer Science and Engineering

Department of Computer Science And Engineering

University of Moratuwa
Sri Lanka

Oct 2018

Sinhala - Tamil Statistical Machine Translation (SMT) for Official Documents

Farook Fathima Farhath

(168034C)

Thesis submitted in partial fulfillment of the requirements for the
Degree of Master of Philosophy in Computer Science and Engineering

Department of Computer Science And Engineering

University of Moratuwa
Sri Lanka

Oct 2018

Declaration

I, Farook Fathima Farhath, declare that this is my own work and this thesis 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 MPhil thesis under my supervision.

Name of Supervisor: Prof. Sanath Jayasena

Signature of supervisor: Date:

Name of Supervisor: Dr. Surangika Ranathunga

Signature of supervisor: Date:

Abstract

Sinhala and Tamil are declared to be the official languages of Sri Lanka. This requires each government related dissemination/communication to be done in both the languages. Even though the requirement for translation is higher, the number of available human translators is limited. One feasible option to boost the productivity would be assisting the human translators with machine translation output. Here the machine translation output is given to translators to work on by post editing, rather than translating from the scratch. However, Sinhala - Tamil pair does not have any well-performing machine translation system. Therefore, the focus of this research is to develop a machine translation system for short official government documents.

This thesis presents two main contributions towards building 'Si-Ta', the first domain-adapted machine translation system for Sinhala - Tamil. The first contribution is building the baseline translation system. The second is implementing data pre-processing techniques to improve the translation quality of the baseline system.

The baseline system was built using Moses, a phrase-based statistical translation system. This was the feasible option with the available resources.

To improve the quality of the translation, three main approaches were explored. They are: (a) domain adaptation, (b) integration of terminology, dictionary, and name lists, and (c) addressing out-of-vocabulary (OOV) problem using word-embedding-based paraphrasing.

In order to adapt the system for the domain of official government documents, different language model design techniques and a data filtration technique were experimented. Under terminology integration, experiments were carried out to evaluate the effect of incorporating bilingual terminology lists to the system. Moreover, a novel data augmentation technique was experimented to generate parallel data using bilingual lists and available parallel data. Further, open domain dictionary entries, as well as a list of person names and addresses were integrated and evaluated. In addition, word-embedding-based paraphrasing was used along with a novel heuristic-based filtering to address the out-of-vocabulary issue.

All the above-mentioned approaches gave an improvement over the baseline, apart from data filtering technique. Yet, all these scores were above the scores of already available machine translation systems for this language pair. Though our techniques/approaches were evaluated only on Sinhala - Tamil pair, they are feasible to be applied to other low-resourced, highly inflectional language pairs.

Keywords: Machine Translation, Sinhala, Tamil, Domain Adaptation, Terminology Integration, Out-of-vocabulary

Acknowledgements

First and foremost, I would like to extend my heartfelt gratitude to my supervisors, Dr. Surangika Ranathunga and Prof. Sanath Jayasena for their great guidance, mentoring, suggestions and their patience over past 2 ½ years. Their worthy feedback and comments helped me a lot in carrying out this research in a successful manner.

Beside my supervisors, I would like to extend my thanks to the members of the NLP Center of the University of Moratuwa. Especially, Dr. Uthayasanker Thayasivam and Prof. Gihan Dias for the assistance and guidance and worthy feedback that helped me with this research. And would much appreciate the precious times spent with my research colleagues.

I would like to acknowledge various bodies, funded my research. Especially, the Senate Research Council (SRC) Grant of the University of Moratuwa and the Department of Official Language, Sri Lanka.

Also, I would extend my gratitude to the Head of the Department and the staff of Department of Computer Science and Engineering for the various supports rendered by them and for giving me the opportunity to pursue this degree.

Further, I am in much debt to my friends who lend a helping hand in various means. Especially their motivating words were a comfort at the times when I needed sympathy.

Finally, I would like to extend thanks to the warriors behind the screen, my husband Azaam, son Rashaadh, my parents and siblings for their great support, commitment, encouragement, and dedications which made this dream a reality. They have dedicated their own time from their busy schedules to support me. And especially I regret and thank Azaam and Rashaadh for being patient at busy times.

Contents

Declaration of Authorship **i**

Abstract **ii**

Acknowledgements **iii**

List of Figures **vii**

List of Tables **ix**

List of Abbreviations **x**

1 Introduction **1**

1.1 Problem Definition 1
1.2 Motivation 2
1.3 Objective of Research 3
1.4 Research Contributions 3
1.5 Publications 3
1.6 Outline 4

2 Literature Review **5**

2.1 Machine Translation History 5
2.2 Different Approaches 7
2.2.1 Rule-Based Machine Translation (RBMT) 7
2.2.1.1 Direct Machine Translation (DMT) 8
2.2.1.2 Transfer-Based Machine Translation Approach (TBMT) 8
2.2.1.3 Interlingual Machine Translation Approach 8
2.2.1.4 Challenges in Rule-Based Systems 9
2.2.2 Corpus-Based Machine Translation Approach 9
2.2.2.1 Example-Based Machine Translation (EBMT) . . . 9
2.2.2.2 Statistical Machine Translation (SMT) 10

2.2.2.3 Neural Machine Translation (NMT) 11
2.2.3 Hybrid Approach 11
2.3 Statistical Machine Translation 12
2.3.1 Translation Model 14
2.3.1.1 Word-Based SMT 14
2.3.1.2 Phrase-Based SMT 17
2.3.2 Language Model 18
2.3.3 Reordering Model 20
2.3.4 Word Penalty 20
2.3.5 Phrase Penalty 21
2.3.6 Decoding (Search) 21
2.3.7 Tuning 23
2.3.8 Evaluation 24
2.3.9 Common Challenges in SMT 26
2.3.10 Prior work in Sinhala-Tamil Translation 27
2.3.11 Challenges in Sinhala-Tamil Machine Translation 29
2.4 Techniques for Improving Translation Quality 31
2.4.1 Domain Adaptation 31
2.4.2 Terminology Integration 33
2.4.2.1 Static Integration 36
2.4.2.2 Dynamic Integration 39
2.4.3 Handling Out-of-Vocabulary Words 40
2.5 Summary 42

3 The Baseline Si-Ta SMT System 44
3.1 Introduction 44
3.2 Selection of Translation Methodology 45
3.3 Data Source Description 45
3.4 Selection of an SMT Framework 48
3.5 Baseline System Setup 48
3.6 Summary 50

4 Techniques for Improving SMT 51
4.1 Introduction 51
4.2 Domain Adaptation in SMT 52
4.2.1 Multiple Models 52
4.2.2 Data Filtering 53
4.3 Terminology Integration in SMT 55
4.3.1 Static Integration Techniques 57
4.3.1.1 As Corpus 57
4.3.1.2 Multiple Tables 58
4.3.1.3 Merged Tables 59

4.3.1.4 Parallel Data Augmentation with Bilingual Lists . . 61
4.3.2 Dynamic Integration Technique 71
4.4 Dictionary Integration 72
4.5 Name List Integration 73
4.6 Handling OOV in SMT 74
4.7 Summary 77

5 Evaluation and Analysis **79**
5.1 Introduction 79
5.2 Baseline 79
5.3 Domain Adaptation 83
5.4 Terminology Integration 85
5.5 Dictionary Integration 90
5.6 Name List Integration 91
5.7 Handling Out-of-Vocabulary Words 92
5.8 Summary 95

6 Conclusion and Future Work **96**

Bibliography **100**

List of Figures

2.1 Vauquois Triangle (source: [1]) 7

2.2 Sample word alignment between Sinhala and Tamil 15

2.3 Possible Tamil translation options for a sample Sinhala sentence . . . 22

2.4 Sample diagram of the decoding process 23

2.5 Conceptual design of static integration of terminology 36

2.6 Conceptual design of dynamic integration of terminology 37

3.1 A sample screenshot where there are mismatches in the punctuation in parallel sentences. 47

3.2 Baseline Si-Ta system 49

4.1 Language models are log-linearly interpolated. LM 1, LM 2 and LM 3 are the individual language models created from the *in-domain, pseudo-in-domain* and *out-domain* data. 54

4.2 Language models are linearly interpolated. LM 1, LM 2 and LM 3 are the Language models created out of the *in-domain, pseudo-in-domain* and *out-domain* data while the LM is the interpolated language model. 55

4.3 Conceptual design of terminology integration-'as corpus' 58

4.4 Sinhala word 'ස්ථිර' having translation variations based on the context and inflection as highlighted. 59

4.5 Conceptual design of terminology integration technique - 'Multiple tables' 60

4.6 Sample screenshot of a phrase table for 'merge-table'. Values for the new feature are highlighted. 61

4.7 Comparison between similar word lists for sample Sinhala words based on Word2vec and fast-text based models. The left side list is fetched from the Word2vec model while the right side one is from the fastText. 63

4.8 Comparison between similar word lists for sample Tamil words based on Word2vec and fast-text based models. The left side list is fetched from the Word2vec model while the right side one is from the fastText. 64

4.9 A sample cluster of terms based on the ending word 65

4.10 A sample similarity list retrieved based on word embedding 66

4.11 An original (first pair) and the augmented sentence pair (second pair) based on *'Based on ending word'* technique. The term replaced is highlighted. 66

4.12 Boundary detection based on *'Based on ending word + POS'* for the same example illustrated in 4.11. 68
4.13 Example to illustrate the difference in *'Based on ending word + improved POS'* and *'Based on ending word + POS'*. The sentence what is being used in the *'Based on ending word + POS'* technique does not meet the requirement of *'Based on ending word + improved POS'* technique since the POS of the word that follows is 'POST'. 70
4.14 A sample sentence where the identification of the term is based on the NE tag. Term boundary is highlighted. 71
4.15 Sample source text, XML pre-processed based on bilingual translation option. 72
4.16 Sample dictionary entries that have multiple translation equivalents . . 73
4.17 Anomalies in the source side of the dictionary entry that need further pre-processing before utilizing it in the parallel corpus. 73
4.18 An example original sentence and an augmented sentence. The paraphrase and the OOV are highlighted. 76

5.1 Five point Likert scale used to evaluate the translation outputs of Si-Ta system and Google translate. 80
5.2 An example where Google Translate gives a misleading translation. . 83
5.3 Alignment for the Sinhala word 'කොන්ක්‍රීට්' /concrete/ (concrete) in the baseline system (Although the equivalent Tamil word '**கான்கிரீட்**' /concrete/ exists, it is misaligned). 86
5.4 Alignment information for the same example as in figure 5.3, for 'as corpus' setup. Here the equivalent Tamil word is aligned. 87

List of Tables

3.1 Statistics on parallel data (S- Sinhala, T- Tamil) 47
3.2 Statistics on Sinhala monolingual data 47
3.3 Statistics on Tamil monolingual data 48

4.1 Perplexity values for different domain data - Sinhala 56
4.2 Perplexity values for different domain data - Tamil 56
4.3 Statistics on the terminology list utilized 57

5.1 Evaluation scores for the baseline system 80
5.2 Time taken (in minutes) for translation of text manually and by post-editing the Si-Ta output for each translator 81
5.3 Translation score variations for different language model configurations. 83
5.4 BLEU scores for different term integration techniques for bilingual term integration (the higher the better). 86
5.5 Number of OOVs for different term integration techniques for bilingual term integration (the lower the better) 86
5.6 Evaluation scores for integration of augmented parallel data generated based on different techniques. 86
5.7 Evaluation scores for (open domain) dictionary integration 90
5.8 Evaluation scores for integration of a bilingual name list 91
5.9 BLEU scores for word embedding-based paraphrasing over OOVs . . 93
5.10 OOV counts for use of word embedding-based paraphrasing over OOVs 94