

# 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 Mokbanarangan

(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 Mokbanarangan, 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

<b>Declaration of Authorship</b>	<b>i</b>
<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>Abbreviations</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Overview of Named Entity Recognition . . . . .	1
1.2 Overview of Machine Translation . . . . .	3
1.3 Motivation . . . . .	4
1.4 Research Objectives . . . . .	4
1.5 Contributions . . . . .	5
1.6 Articles . . . . .	5
1.7 Organization of the Thesis . . . . .	6
<b>2 Literature Review</b>	<b>7</b>
2.1 Named Entity Recognition . . . . .	8
2.1.1 Challenges in NER . . . . .	8
2.1.2 Datasets . . . . .	9
2.2 Existing Approaches for Named Entity Recognition . . . . .	10
2.2.1 Rule-based Approaches . . . . .	11
2.2.2 Machine Learning Approaches . . . . .	11
2.2.3 Semi-supervised Approaches . . . . .	15
2.2.4 Unsupervised Approaches . . . . .	16
2.2.5 Cross Lingual Approaches . . . . .	17

2.2.6	Deep-Learning Approaches . . . . .	18
2.2.7	Existing Approaches Used for Tamil and Sinhala NER . .	19
2.2.7.1	Tamil . . . . .	19
2.2.7.2	Sinhala . . . . .	20
2.2.8	Existing Approaches Used in Different South Asian Lan- guages . . . . .	20
2.2.9	Features in NER . . . . .	21
2.2.9.1	Local Features . . . . .	21
2.2.9.2	Global Features . . . . .	22
2.2.9.3	Resources . . . . .	23
2.2.10	Available Platforms and Toolkits . . . . .	23
2.2.10.1	Stanford NER . . . . .	23
2.2.10.2	GATE Named Entity Recognizer . . . . .	24
2.2.10.3	Natural language Toolkit (NLTK) . . . . .	24
2.2.11	Evaluation Measures . . . . .	24
2.2.12	Summary . . . . .	25
2.3	Graph Based Semi-Supervised Learning (GSSL) . . . . .	26
2.3.1	Graph-based Approach for Sequential Tagging . . . . .	27
2.3.2	Summary . . . . .	29
2.4	Distributional Semantic Models - DSM . . . . .	29
2.4.1	Pointwise Mutual Information (PMI) Vector . . . . .	30
2.4.2	Word2Vec . . . . .	31
2.4.3	FastText . . . . .	31
2.4.4	Wang2Vec . . . . .	32
2.4.5	ELMo . . . . .	32
2.4.6	Summary . . . . .	33
2.5	Machine Translation . . . . .	33
2.5.1	Statistical Machine Translation - SMT . . . . .	34
2.5.1.1	Moses . . . . .	35
2.5.2	Neural Machine Translation - NMT . . . . .	36
2.5.2.1	Encoder-Decoder Model . . . . .	37
2.5.3	Evaluation . . . . .	37
2.5.3.1	BLEU Score . . . . .	37
2.5.3.2	NIST Score . . . . .	38
2.5.4	Existing Machine Translation Systems for Tamil-to-Sinhala Translation . . . . .	39
2.5.4.1	<i>SiTa</i> SMT system . . . . .	41
2.5.5	Existing Approaches to Translate Named Entities . . . . .	41
2.5.6	Summary . . . . .	43
<b>3</b>	<b>Methodology</b> . . . . .	<b>44</b>
3.1	Identifying the Tag Set . . . . .	46



---

3.2	Annotated Dataset . . . . .	47
3.3	Building the Named Entity Recognizer for Tamil and Sinhala . .	49
3.3.1	Graph Based Semi-Supervised Learning . . . . .	49
3.3.1.1	Representing Nodes of Graph . . . . .	49
3.3.1.2	Creating Edges of the Graph . . . . .	50
3.3.1.3	Label Propagation . . . . .	52
3.3.2	Bi-directional LSTM CRF Sequential Tagging . . . . .	52
3.3.2.1	Character Embedding . . . . .	54
3.3.2.2	Predicting the Tags . . . . .	56
3.3.2.3	Tuning the Hyper-parameters . . . . .	57
3.4	Translating Identified Named Entities . . . . .	58
3.4.1	Unsupervised Morphology Induction . . . . .	59
3.4.2	Integrating to Moses . . . . .	60
<b>4</b>	<b>Implementation</b>	<b>62</b>
4.1	Building the Corpus . . . . .	62
4.2	Building the Named Entity Recognizer . . . . .	62
4.2.1	AllenNLP Research Library . . . . .	62
4.2.2	Building the Word embedding Models . . . . .	64
4.2.3	Modifying the metric-learn library . . . . .	65
4.2.4	Implementing Graph Based Semi-supervised Sequential Tag- ging Algorithm . . . . .	66
4.2.5	Implementing BiLSTM CRF Tagging . . . . .	67
4.3	Integrating to Moses . . . . .	67
4.3.1	SiTa System . . . . .	67
<b>5</b>	<b>Experiments and Results</b>	<b>69</b>
5.1	Graph Based Semi Supervised Learning . . . . .	70
5.2	Bi-directional LSTM CRF Tagging . . . . .	76
5.2.1	NER . . . . .	76
5.2.2	POS . . . . .	77
5.3	Integrating to Moses . . . . .	79
5.3.1	Sinhala → Tamil Translation . . . . .	80
5.3.2	Tamil → Sinhala Translation . . . . .	80
<b>6</b>	<b>Conclusion</b>	<b>82</b>
<b>7</b>	<b>Future Works</b>	<b>84</b>
	<b>Bibliography</b>	<b>104</b>

## List of Figures

1.1	An example of NER application on an example text . . . . .	2
2.1	CBOW Vs Skip-gram models . . . . .	32
2.2	Neural Machine Translation . . . . .	37
3.1	Outline to build the NER . . . . .	46
3.2	Named entity tag distribution for Sinhala . . . . .	48
3.3	Named entity tag distribution for Tamil . . . . .	49
3.4	A bidirectional LSTM network . . . . .	53
3.5	A BiLSTM-CRF model . . . . .	54
3.6	Architecture of the BiLSTM network with a CRF Classifier . . .	55
3.7	Character-based representation using convolutional neural network	56
3.8	Character-based representation using BiLSTM networks . . . . .	57
3.9	Preprocessing the input data . . . . .	61
4.1	The Tagtog annotation tool . . . . .	63
5.1	English POS accuracy for GSSL Vs LSTM-CRF . . . . .	74
5.2	English chunking F1-Score for GSSL Vs LSTM-CRF . . . . .	75
5.3	English NER F1-Score for GSSL Vs LSTM-CRF . . . . .	75

## List of Tables

2.1	Different approaches for NER in Indian languages . . . . .	21
2.2	SMT and GIZA++ based approaches for Sinhala-Tamil Translation	40
3.1	NER Tag set . . . . .	47
3.2	Corpus Kappa scores . . . . .	48
4.1	Perplexity scores for ELMo Model . . . . .	64
5.1	Comparison of different methods to represent nodes and their respective accuracy for different tasks in English. A - Single Vector, B - Dimension reduced Single Vector, C - Concatenated $n$ -gram vectors, D - Dimension reduced concatenated $n$ -gram vectors. . .	72
5.2	Comparison of different methods to represent nodes and their respective accuracy for Tamil and Sinhala POS tagging. A - Single Vector, B - Dimension reduced Single Vector, C - Concatenated $n$ -gram vectors, D - Dimension reduced concatenated $n$ -gram vectors.	73
5.3	Comparison of different methods to represent nodes and their respective F1-scores for Tamil and Sinhala NER tagging. A - Single Vector, B - Dimension reduced Single Vector, C - Concatenated $n$ -gram vectors, D - Dimension reduced concatenated $n$ -gram vectors.	74
5.4	Comparison of different vectors and their respective accuracy for Tamil and Sinhala NER tagging with BiLSTM CRF. A - FastText, B - Wang2Vec, C - ELMo, D - ELMo + Wang2Vec, E - ELMo + FastText . . . . .	76
5.5	Comparison of different vectors and their respective accuracy for Sinhala POS tagging with BiLSTM CRF. A - FastText, B - Wang2Vec, C - ELMo, D - ELMo + Wang2Vec, E - ELMo + FastText . . . .	77
5.6	Comparison of different vectors and their respective accuracy for Tamil POS tagging with BiLSTM CRF . . . . .	78
5.7	SMT integration experiments . . . . .	79
5.8	Sinhala→Tamil translation scores after named entity translation integration . . . . .	80
5.9	Tamil→Sinhala translation scores after named entity translation integration . . . . .	80

## Abbreviations

<b>NLP</b>	Natural Langaguge <b>P</b> rocessing
<b>NER</b>	Named <b>E</b> ntity <b>R</b> ecognition
<b>NE</b>	Named <b>E</b> ntities
<b>BLEU</b>	Bi-Lingual <b>E</b> valuation <b>U</b> nderstudy
<b>MT</b>	Machine <b>T</b> ranslation
<b>CRF</b>	Conditional <b>R</b> andom <b>F</b> ield
<b>LSTM</b>	Long <b>S</b> hort <b>T</b> erm <b>M</b> emory
<b>GSSL</b>	Graph <b>B</b> ased <b>S</b> emi- <b>S</b> upervised <b>L</b> earning
<b>CBOW</b>	Continous <b>B</b> ag <b>O</b> f <b>W</b> ords
<b>ME</b>	Maximum <b>E</b> ntropy
<b>HMM</b>	Hidden <b>M</b> arkov <b>M</b> odel
<b>POS</b>	Parts <b>O</b> f <b>S</b> peech
<b>SMT</b>	Statistical <b>M</b> achine <b>T</b> ranslation
<b>NMT</b>	Neural <b>M</b> achine <b>T</b> ranslation