

**USING MULTI AGENT TECHNOLOGY FOR
AUTOMATIC MACHINE TRANSLATION**

Budditha Hettige

(118036M)

Degree of Doctor of Philosophy

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

July 2020

USING MULTI AGENT TECHNOLOGY FOR AUTOMATIC MACHINE TRANSLATION

Budditha Hettige

(118036M)

Thesis submitted in partial fulfillment of the requirements for the degree
Doctor of Philosophy

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

July 2020

Declaration

I 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 the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or part in print, electronic or another medium. I retain the right to use this content in whole or part in future works (such as articles or books)



13.07.2020

.....

.....

Signature:

Date:

Budditha Hettige

Candidate

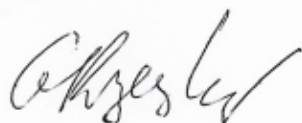
The above candidate has carried out the research for the PhD thesis under my supervision.

.....

.....

Prof. Asoka S. Karunananda

Date:



14.07.2020

.....
Prof. George Rzevski

.....
Date:

Dedicated to

This thesis is dedicated

...to my beloved mother and father

...to my wife and son

Acknowledgement

Many people have helped their best to successfully completion of this research. I acknowledge all of them for their valuable thoughts, and constant encouragement gave me to make my research a reality.

First and foremost, I acknowledge my supervisor senior professor Asoka Karunananda for accepting me as his research student and giving excellent support and advice. Prof. Karunananda is a great mentor who guided while giving me all the freedom and encouragement to accompany with my ideas. Also, I acknowledge my second supervisor, professor George Rzevski for accepting me as his research student and giving excellent support and advice. Without both them patient listening and creative thoughts, this work would not have been possible at all.

Especially, I acknowledge Venerable Kirioruwe Dhammananda Thero for his kind-hearted help and encouragement to fulfil my research work and give correct direction to my successful life.

My very special thank goes to Dr (Mrs.) Uditha Rathnayake, Dr (Mrs.) Menaka Ranasinghe and Dr Lochandraka Ranathunga for their invaluable comments and guidance as my examiners of bi-annual review panels.

Also, I wish to extend my sincere thanks for the support I received from all the members of the administration office and members of the Faculty of Information Technologies, University of Moratuwa. Especially I thank Dr (Mrs.) Thushari Silva and Dr (Mrs.) Subha Fernando and Madam, Prof. Deelika Dias for their essential roles.

I also acknowledge Mr P. Dias (Department of Statistics, University of Sri Jayewardenepura) and Ms. E.R.C. Sadamali for their kind support to fulfil my research work.

Also, exceptional and heartfelt thanks for Dr. (Mrs.) Mihirini Wagaarachchi, Dr (Mrs.) Chinthani Weerakoon, Dr. (Mrs.) Anuradha Ariyarthne and Ms. Mihiri Serisooriya for their gracious associations throughout the last couple of years.

I would like to thankfully remind all the academic and non-academic members of the General Sir John Kotelawala Defence University those who have supported me during the work. Special thanks go to the Vice-chancellor Major General Milinda Peiris, the Dean, Faculty of Computing, Commodore J. U. Gunaseela, for granting me the study leave to complete my research.

Again, I would like to mention Dr Asele Gunasekara, Maj R.M. Rathnayaka and all the members of my faculty gave me a great support

Finally, I would like to extend my greatest gratitude to my family members, especially my wife Lakshmi and my little hart Tenuja for the unrestricted support given, and without their care, this would have been unmanageable. Again, I must give an express thanks to Lakshmi for tolerating my busy schedules due to the research work. Last but not least, I thank all who supported me to make this work a success.

July 13, 2020

Budditha Hettige

Abstract

Machine translation is a cost-effective, quick, and widely accepted automated language translation method that has become essential in the modern and ever more globalized world. Machine translation can be done with one or more different approaches, including dictionary-based, rule-based, example-based, phrase-based, statistical, or neural-linguistic approaches. Nevertheless, most of the existing machine translation systems show a quality gap when compared with human translation. Thus, human translation has been considered as the best language translation method so far. Human language translation is a complex and opportunistic process depends on human memory. This human language translation process has been described through a few theories. Among them, the garden path model and the constraint satisfaction model are two fundamental approaches available for human language translation, especially concerning sentence parsing with meaning. These two theoretical models demonstrate how to select suitable words in the phrase of a sentence to generate accepted meanings. Based on these two theories, a hybrid approach to machine translation has been proposed. This proposed approach is stimulated by how people parse and translate a sentence by putting available phrases together with accepted meaning. According to the approach, translation is done in three stages. In the first stage, the system analyses the given sentence by considering the morphology, syntax, and semantics of the source language. Then, the system uses phrase-based translation and translates each phrase into the target with multiple solutions. The phrase translation is done considering the four factors of psycholinguistic parsing techniques, such as phrase structure, semantic features, thematic roles, and probability. Finally, considering all the translated phrases, the system should be capable of identifying suitable target language phrases to take accepted meanings, considering subject-verb and object-verb agreements. After the subject-verb-object agreement, other available phrases in the sentence should be capable of re-arranging according to the accepted subject, object, and verb phrases.

This approach has been simulated with the multi-agent system named EnSiMaS, which translates English text into Sinhala. The EnSiMaS was implemented on the MaSMT framework, which was specially developed for agent-based machine translation. The EnSiMaS comprises of 26 language processing agents on both source and target languages. These agents were clustered into six agent swarms considering morphological, syntactical, and semantical concerns of the source and the target languages. In addition to these language-processing agents, the system should be able to create an agent dynamically for each source language phrase. These dynamically created phrase agents should be capable of communicating with other relevant phrases and taking the accepted solutions.

The EnSiMaS was tested with 85 sample English sentences. For each English sentence, three different translations were taken. According to the evaluation result, the system shows an 8.77% word error rate, a 6.72% inflexion error rate, and a 5.37% sentence error rate for the first, second, and third translations. In addition, calculated BLUE scores show 0.89160756, 0.52009204, and 0.43581893 for the first, second, and third translations. Then randomly selected 25 samples sentences are used to calculate the adequacy and fluency of the EnSiMaS. Adequacy and fluency rates were taken from 55 human evaluators considering the human-translated reference sentences. The Kendal's Tau correlation coefficient shows that there is a weak positive association between adequacy levels of human translations vs EnSiMaS system translations and moderate positive association between fluency levels of human translation and EnSiMaS system translation. Further, according to the Fleiss Kappa coefficient method, there is a significant fair agreement on raters for adequacy and fluency ratings.

Keywords: Machine Translation, Multi-agent systems, Human Language Processing, MaSMT, EnSiMaS

Table of Contents

Declaration	i
Acknowledgement	iii
Abstract	v
Table of Contents	vi
List of Figures	xiii
List of Tables	xvi
List of Abbreviations	xviii
CHAPTER 1	1
INTRODUCTION	1
1.1 Prolegomena	1
1.2 Aim and Objectives	3
1.3 Problem in Brief	4
1.4 The Scope of the Research	4
1.5 Hypothesis	4
1.6 Human Translation to Machine Translation	5
1.7 Proposed Approach to Machine Translation	6
1.8 Resource Requirements	7
1.9 Chapter Organisation	8
1.10 Summary	9
CHAPTER 2	10
STATE OF THE ART IN MACHINE TRANSLATION	10
2.1 Introduction	10
2.2 Fundamentals of Machine Translation	10
2.3 Brief History	12
2.4 Existing Approaches to Machine Translation	14
2.4.1 Interlingua approach	14
2.4.2 Human-Assisted/Computer-Aided Translation	16
2.4.3 Dictionary-based Machine Translation	18
2.4.4 Rule-based Machine Translation	19
2.4.5 Example-based Machine Translation	21

2.4.6 Statistical Approach to Machine Translation	22
2.4.7 Neural Machine Translation	24
2.4.8 Knowledge-based Approach	27
2.4.9 Transfer-based Machine Translation	27
2.4.10 Agent-based Approach to Machine Translation	28
2.4.11 Hybrid Approach to Machine Translation	29
2.5 Local Resource and Existing ESMTS	30
2.6 Some Issues in Machine Translation	31
2.6.1 Word and Sentence Segmentation	31
2.6.2 Word Conjugation	31
2.6.3 Tense Detection	32
2.6.4 Multi-word Expression	32
2.6.5 Out of Vocabulary	32
2.6.6 Translating Idiomatic Phrases	32
2.7 Summarization of Existing MT Approaches	33
2.8 Problem Definition	34
2.9 Summary	34
CHAPTER 3	35
LITERATURE REVIEW AND BACKGROUND	35
3.1 Introduction	35
3.2 Computational Grammar for the English Language	35
3.2.1 The Morphology of the English Language	35
3.2.2 Syntax of the English Language	40
3.2.7 The Semantics of English Language	42
3.3 The Sinhala Language	44
3.3.1 Morphology of the Sinhala Language	44
3.3.2 Syntax of the Sinhala Language	48
3.4 Comparison Between English and Sinhala Languages	49
3.5 Summary	49

CHAPTER 4	50
NATURAL LANGUAGE PROCESSING TECHNIQUES	50
4.1 Introduction	50
4.2 Computational Model for English and Sinhala	50
4.3 Morphological Analysis and Generation	53
4.4 Syntactical Analysis and Generation	55
4.5 Semantics Processing	56
4.5.1 Word level semantics	56
4.5.2 Phrase level semantics	57
4.5.3 Sentence level semantics	57
4.6 Summary	57
CHAPTER 5	58
MULTI AGENT TECHNOLOGY	58
5.1 Introduction	58
5.2 What is Multi-agent System?	58
5.2.1 Type of Agents	59
5.2.2 Agent Communication	60
5.3 Existing MAS Development Framework	60
5.4 MaSMT: Multi-agent Framework for Machine Translation	63
5.4.1 MaSMT Framework	64
5.4.2 AGR Organisational Model and MaSMT Architecture	64
5.4.3 MaSMTAbstractAgent	66
5.3.4 MaSMT Agent	66
5.3.5 MaSMT Agent’s Life cycle	67
5.3.6 MaSMT Controller agent	67
5.3.7 MaSMT Root Agent	67
5.3.8 MaSMT Agents’ Swarm	68
5.3.9 MaSMT Messages	69
5.3.10 MaSMT Settings	70
5.3.11 MaSMT Message Parsing	71
5.3.12 Applications of MaSMT	72

5.4 Summary	72
CHAPTER 6	73
A HYBRID APPROACH TO MACHINE TRANSLATION	73
6.1 Introduction	73
6.2 A Novel Approach to Machine Translation	73
6.3 Theoretical Basis of Language Translation	74
6.4 A multi-agent Approach to Machine Translation	76
6.4.1 Multi-agent Approach to English Morphological Analysis	76
6.4.2 Multi-agent Approach to English Syntax Analysis	76
6.4.3 Multi-agent Approach to English to Sinhala Phrase-based Translation	77
6.4.4 Multi-agent Approach to Sinhala Morphological Generation	77
6.4.5 Multi-agent Approach to Sinhala Syntax Generation	78
6.5 Why a Multi-agent Approach?	78
6.6 Features of EnSiMaS	79
6.7 Input for EnSiMaS	79
6.8 Output of EnSiMaS	79
6.9 Process of the EnSiMaS	80
6.10 Summary	81
CHAPTER 7	82
DESIGN OF THE ENSIMAS	82
7.1 Introduction	82
7.2 Design of the EnSiMaS	82
7.2.1 EnSiMaS GUI	83
7.2.2 Ontology	83
7.2.3 Virtual World	84
7.2.4 English Morphological Swarm	85
7.2.5 English Syntax Analysing Swarm	86
7.2.6 Bilingual Semantics Swarm	87
7.2.7 Sinhala Morphological Swarm	88
7.2.8 Sinhala Syntactical Swarm	89
7.2.9 Ontological Swarm	89

7.2.10 English Phrase-based Translation Swarm	90
7.3 Summary	91
CHAPTER 8	92
IMPLEMENTATION OF ENSIMAS	92
8.1 Introduction	92
8.2 EnSiMaS Ontology	92
8.2.1 English Pronoun Table	93
8.2.2 English Regular Noun table	94
8.2.3 English Regular Verb Table	94
8.2.4 English Irregular noun table	95
8.2.5 English Irregular verb table	95
8.2.6 English Regular Adjective Table	96
8.2.7 Other word table	96
8.2.8 English-Sinhala Bilingual Dictionary	97
8.2.9 Morphological rules for English words	97
8.2.10 Syntax Rule for English phrases	98
8.3 EnSiMaS Virtual World	99
8.3.1 The English Word and Word List	100
The English Word Morphology	101
8.3.2 The English Phrase and Phrase List	101
8.3.3 The Sinhala Phrase and Sinhala phrase list	102
8.3.4 The Sinhala word Lexicon	103
8.3.5 The EnSiMaS Phrase	104
8.3.6 The EnSiMaS Phrase List	105
8.3.7 The EnSiMaS Sentence info	106
8.4 EnSiMaS Agents	106
8.4.1 EnSiMaS Manager Agent	106
8.4.2 English Morphological System	107
8.4.3 English Syntax Swarm	110
8.4.4 Bilingual Semantic Swarm	113
8.4.5 Sinhala Morphological Generation Swarm	115

8.4.6 Sinhala Syntactical Swarm	117
8.4.7 Translation Controller Agent	117
8.4.8 Translation Swarm	118
8.4.9 Ontological Swarm	119
8.5 Summary	120
CHAPTER 9	121
EVALUATION	121
9.1 Introduction	121
9.2 English to Sinhala Multi-agent System (EnSiMaS)	121
9.3 EnSiMaS Dictionary	122
9.4 EnSiMaS Translator	123
9.5 EnSiMaS Phrase-based Editor	128
9.6 Evaluation Strategy of the EnSiMaS	129
9.6.1 Round Trip Translation	130
9.6.2 Word Error Rate	130
9.6.3 Sentence Error Rate	131
9.6.5 Inflectional Error Rate	132
9.6.6 BLEU	132
9.6.7 METEOR	133
9.6.8 Human Evaluation	133
9.7 Experiment	137
9.8 EnSiMaS vs Google Translator	140
9.9 Results and Data Analysis	141
9.9.1 Details of the sample set	141
9.9.2 Adequacy and Fluency	143
9.10 Conclusion of the Data Analysis	149
9.11 Summary	150
CHAPTER 10	151
CONCLUSION AND FURTHER WORK	151
10.1 Introduction	151
10.2 Hybrid Approach for Machine Translation	151

10.3 Conclusion	152
10.4 Objectives-wise Achievement	154
10.5 Limitations	157
10.6 Further Work	157
10.7 Summary	158
References	159
Appendix A: Translation Summary with Agents' Communications	180
Appendix B: EnSiMaS User Manual	198
Appendix C: Sample of evaluation form	201
Appendix D: List of Publications	204
Appendix E: MaSMT Development Guide	205

List of Figures

Figure 2.1: General pipeline of the MT	10
Figure 2.2: Machine translation pyramid	11
Figure 2.3: Historical Development of the MT	13
Figure 2.4: Taxonomy of the Machine Translation	14
Figure 2.5: Translation Process of the interlingua MT	15
Figure 2.6: General pipeline of the dictionary-based MT system	18
Figure 2.7: Design of the dictionary-based MT	19
Figure 2.8: Components of the RBMT system	20
Figure 2.9: Activities on statistical MT	23
Figure 2.10: Encoder-decoder architecture of the NMT	25
Figure 3.1: Part of speech mapping between English and Sinhala	44
Figure 4.1: Language model for English and Sinhala	51
Figure 4.2: Ontology of a word	51
Figure 4.3: Ontology for a Phrase	52
Figure 4.4: Ontology for a Sentence	53
Figure 4.5: Process of the morphological analysis and generation	55
Figure 5.1: Different types of agents	59
Figure 5.2: UML-Based Aalaadin model for multi-agent system development	64
Figure 5.3: Agents' architecture on MaSMT	65
Figure 5.4: Modular architecture of the MaSMT Agent	66
Figure 5.5: The life cycle of the MaSMT Agent	67
Figure 5.6: Architecture of the MaSMT controller agent	68
Figure 5.7: Design of the swarm of Agents	69
Figure 6.1: Factors contribute to sentence parsing	74
Figure 7.1: Design of the EnSiMaS	82
Figure 7.2: Design of the EnSiMaS Ontology	84
Figure 7.3: Design of the EnSiMaS Virtual world	85
Figure 7.4: Design of the English morphological swarm	86
Figure 7.5: Syntax analyzing swarm: agents for English syntax analysis	87
Figure 7.6: Design of the Bilingual Semantics swarm	88

Figure 7.7: Design of the Sinhala morphological swarm	89
Figure 7.8: Design of the phrase-based translation Swarm	90
Figure 8.1: Structure and sample data on the pronoun table	93
Figure 8.2: Structure and sample data on the regular noun table	94
Figure 8.3: Structure and sample data on the regular verb	94
Figure 8.4: Structure and sample data on the irregular noun	95
Figure 8.5: Structure and sample data on the irregular Verb table	95
Figure 8.6: Structure and sample data on the regular adjective table	96
Figure 8.7: Structure and sample data on the other word table	96
Figure 8.8: Structure and sample data on the bilingual dictionary	97
Figure 8.9: sample data for English morphological rules	98
Figure 8.10: Selected rules to detect English phrases (Phrase rules)	99
Figure 8.11: Class diagrams of the English word and English wordlist	100
Figure 8.12: Class diagrams of the English word morphology	101
Figure 8.13: Class diagrams of the English phrase and English phrase list	102
Figure 8.14: Class diagrams of the Sinhala Phrase and Sinhala phrase list	103
Figure 8.15: Class diagrams of the Sinhala word lexicon and Sinhala word lexicon list of the EnSiMaS	104
Figure 8.16: Class diagram of the EnSiMaS phrase	105
Figure 8.17: Class diagram of the EnSiMaS phrase list	106
Figure 8.18: Activity diagram of the morphological agent	108
Figure 8.19: Communication diagram of the EMS	109
Figure 8.20: Activity diagram of the English Syntax Swarm	112
Figure 8.21: Communication diagram of the syntactical swarm	113
Figure 8.22: Communication diagram of the Bilingual semantics swarm	114
Figure 8.23: Activities of the Sinhala Noun Generation agent	116
Figure 8.24: Agent communication diagram of the translation swarm	119
Figure 9.1: Top-level application selection GUI of the EnSiMaS	121
Figure 9.2: GUI of the EnSiMaS dictionary	122
Figure 9.3: A dictionary-based bilingual word editor	123
Figure 9.4: GUI of the EnSiMaS Translator	128
Figure 9.5: GUI of the EnSiMaS phrase-based editor	129

Figure 9.6: Distribution of the number of words among input sentences	142
Figure 9.7: Percentage distribution on adequacy and fluency values for EnSiMaS Best Translation	145
Figure 9.8: Distribution on adequacy rates on five different raters	146
Figure 9.9: Distribution on fluency rates on five different raters	147

List of Tables

Table 2.1: Summary of the selected MT systems	33
Table 3.1: Some Inflectional suffixes in English	36
Table 3.2 Some Morphological rules for Noun Inflection	37
Table 3.3 Regular and irregular noun forms	37
Table 3.4: Regular and irregular English verb forms	38
Table 3.5: Some Morphological rules for Verb conjugation	38
Table 3.6: Adjective relationship of a noun	39
Table 3.7: Verb and adverb usage	39
Table 3.8: Basic Thematic Relationship in a sentence	43
Table 3.9: Sinhala Noun inflexion form for base word මුඛ (dear)	45
Table 3.10: Verb inflexion forms for Verb <i>maranawa</i> (මරණවා)	46
Table 3.11: Add-remove values for the Sinhala verb	47
Table 3.12: Fundamental differences in both Sinhala and English	49
Table 4.1: Summary of the Existing Morphological analyzers	54
Table 5.1: summary of the existing Multi-agent system development frameworks	63
Table 5.2: Structure of the MaSMT Messages	70
Table 5.3: Default settings of the MaSMT	70
Table 5.4: Message directives (headers for messages)	71
Table 8.1: Statistics of the EnSiMaS Knowledgebase	92
Table 8.2: Agents' details of the English morphological swarm	107
Table 8.3: Morphological Tags	109
Table 8.4: Agents' details of the English syntax swarm	110
Table 8.5: Agents' details of the Bilingual Semantic swarm	114
Table 8.6: Agents' details of the Sinhala morphological swarm	115
Table 8.7: Sinhala Syntax Generation Swarm	117
Table 8.8: Ontological Swarm	120
Table 9.1: 1-5 Scale Adequacy matrix	134
Table 9.2: Fluency value in the Likert scale	134
Table 9.3: Fleiss' Kappa values for agreements	135

Table 9.4: Fleiss' Kappa values for agreements	136
Table 9.5: 25 Sample sentences with translated results	138
Table 9.6: Comparison between EnSiMaS vs Google Translator	139
Table 9.7: Summary of descriptive statistics of the 85 input sentences	141
Table 9.8: Calculated WER, IER and SER for the translations	142
Table 9.9: Calculated BLEU results for each translation	143
Table 9.10: Fleiss' kappa coefficient values for Adequacy	144
Table 9.11: Fleiss' kappa coefficient values for Fluency	145
Table 9.12: Summary of the Kendall's rank correlation coefficient for adequacy between Human translation and EnSiMaS translation	148
Table 9.13: Summary of the Kendall's rank correlation coefficient for fluency between Human translation and Fluency on EnSiMaS Translation	149

List of Abbreviations

AI	- Artificial Intelligence
ARGM	- Agent Role Group Model
BCE	- Before the Current Era
BEES	- Bilingual Expert for English to Sinhala
CE	- Current Era
CAT	- Computer Assisted Translation
CYK	- Cocke–Younger–Kasami
CSM	- Constraint Satisfaction Model
EMA	- English Morphological Analysis
ESA	- English Syntax Analysis
ESMTS	- English to Sinhala Machine Translation System
EnSiMaS	- English to Sinhala Multi-Agent System
EBMT	- Example Based Machine Translation
FIPA	- Foundation for Intelligent Physical Agents
GNMT	- Google’s Neural Machine Translation
GPM	- Garden Path Model
HAMT	- Human-assisted (-aided) machine translation
IER	- Inflexion Error Rate
JADE	- Java Agent DEvelopment Framework
KQML	- Knowledge Query and Manipulation Language
LSTM	- Long Short Term Memory
LL	- Left-to-right, Leftmost derivation
LR	- Left-to-right, Rightmost derivation
MWE	- Multi-Word Expressions
MAS	- Multi-Agent System
MT	- Machine Translation
MaSMT	- Multiagent System for Machine Translation
NMT	- Neural Machine Translation
NLTK	- Natural Language Toolkit
NPMT	- Neural Phrase-based Machine Translation

NLP	- Natural Language Processing
PPO	- Preposition Phrase Order
RBMT	- Rule-based Machine Translation System
RTT	- Round-trip Translation
SL	- Source Language
SMT	- Statistical Machine Translation
SER	- Sentence Error Rate
SMG	- Sinhala Morphological Generation
SSG	- Sinhala Syntax Generation
SOV	- Subject Object Verb
SVO	- Subject Verb Object
SPADE	- Smart Python multi-Agent Development Environment
TAG	- Tree Adjoining Grammar
TL	- Target Language
WER	- Word Error Rate