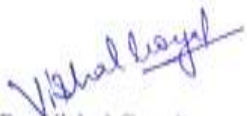


DECLARATION

I hereby affirm that the work presented in this thesis is exclusively my own and there are no collaborators. It does not contain any work for which a degree/diploma has been awarded by any other University/Institution. A part of this thesis has already been published in journals/conferences/workshops proceedings of International and National repute.


(Umrinder Pal Singh)

Countersigned


(Dr. Vishal Goyal)
Associate Professor,
Department of Computer Science,
Punjabi University, Patiala,
(Supervisor)


(Dr. Gurpreet Singh Lehal)
Professor,
Department of Computer Science,
Punjabi University, Patiala,
(Co-Supervisor)

ACKNOWLEDGEMENTS

I would like to manifest my gratitude to my supervisors, Dr. Vishal Goyal, and Dr. Gurpreet Singh Lehal. Their advice, eternal interest, and support made the thesis ultimately possible. Despite their busy schedule, they have been always ready for discussion. I would like to extend my thanks to all those who have helped in concluding my work on time. I will not go into listing all the names here, but I remember even the slightest instance of support provided to me. I am grateful to the teaching and nonteaching staff of Department of Computer Science, Punjabi University, Patiala, for providing help in the systemization of this task. Last, but not the least, I thank all the members of my family for extending me much needed support and motivation to complete this laborious task.



Umrinder Pal Singh

TABLE OF CONTENTS

Certificate	I
Declaration	ii
Abstract	iii
Acknowledgements	iv
Table of Contents	v-vi
List of Tables	vii-viii
List of Figures	lx
List of Charts	X
1. Introduction.....	1-30
1.1 Introduction.....	1
1.2 History Of Machine Translation.....	2
1.3 Approaches To Machine Translation.....	4
1.3.1 Rule Based MT System.....	5
1.3.1.1 Types Of Rule Based MT Systems.....	6
1.3.1.1.1 Direct Mt Systems.....	6
1.3.1.1.2 Transfer Based Systems.....	8
1.3.1.1.3 Interlingual MT Systems.....	10
1.3.2 Corpus-Based Approach	13
1.3.2.1 Statistical MT System.....	13
1.3.3.2 Example Based Machine Translation.....	18
1.3.4 Hybrid Mt System.....	21
1.4 Key Components Of Rule-Based And Corpus-Based Approach	23
1.5 Research Questions.....	26
1.5.1 Challenges To Develop Urdu To Punjabi Mt System.....	26
1.5.1.1 Resource Poor Languages.....	26
1.5.1.2 Spelling Variation.....	27
1.5.1.3 Free Word Order.....	27
1.5.1.4 Segmentation Issues In Urdu.....	28
1.5.1.5 Morphological Rich Languages.....	29
1.5.1.6 Word Without Diacritical Marks.....	30
1.6 Objectives.....	30
2. About Urdu And Punjabi.....	31-44
2.1 Urdu	32
2.2 Punjabi.....	34
2.3 Urdu-Punjabi Grammar.....	36
2.4 About Urdu, Punjabi And Relation With Hindustani And Other Indian Languages	44
3. Review Of Literature.....	45-76
4. Methodology.....	77-111
4.1 Introduction.....	77
4.2 Tokenization And Segmentation Process.....	78

4.2.1	Tokenization Into Sentences.....	79
4.2.2	Tokenization Into Words.....	79
4.2.3	Segmentation Process.....	79
4.3	Text Classification.....	86
4.4	Training Model.....	89
4.4.1	Translation Model Training.....	90
4.4.2	Language Model Training.....	94
4.5	Decoding Model.....	96
4.6	Sub-Modules To Handle Unknown Words.....	96
4.6.1	Remove Diacritical Marks.....	98
4.6.2	Checks For Izafaats.....	99
4.6.3	Stemming.....	101
4.6.4	Creating Inflections.....	104
4.5.5	Transliteration.....	105
4.5.5.1	Issues In Punjabi To Urdu Transliteration.....	105
4.7	Complete Working Algorithm.....	111
5.	Result And Discussion.....	112-124
5.1	Introduction.....	112
5.2	Training And Testing Of Baseline System.....	112
5.3	Bleu Score.....	112
5.4	Manual Testing.....	115
5.5	Evaluation Of Text Classifier	116
5.6	Results With And Without Text Classification.....	118
5.7	Results With And Without Handling Unknown Words.....	119
6.	Conclusion And Future Scope.....	125-126
References	127-134
Appendix	A - Training Files.....	135-139
Appendix	B - Graphical User Interface Of Urdu To Punjabi Mt System..	140-144
Appendix	C - Output Comparison Of All Domains.....	145-147
Appendix	D - List Of Stop Words Used In Text Classification.....	148-149
Appendix	E - Media Coverage.....	150-151
Publication	152

LIST OF FIGURES

Fig.No	Figure Title	Page No.
1.1	Translation pyramid of Bernard Vauquis	5
1.2	<i>Abstract Model of Rule Based MT System</i>	8
1.3	Transfer Based MT System	10
1.4	Interlingua Machine Translation System	11
1.5	Segmented of "Pyramid of Bernard Vauquis"	12
1.6	<i>Abstract Model of Statistical MT System</i>	13
1.7	Example Based MT System	19
1.8	<i>Abstract Model of Hybrid MT System</i>	22
4.1	Incremental MT training and decoding system	78
4.2	Text classification system	87
4.3	Statistical machine translation model	89
4.4	Sub-modules to handle unknown word	98
4.5	Inter module processes to handle unknown words	110

LIST OF CHARTS

Chart No.	Chart Title	Page No.
5.1	Political News Accuracy	113
5.2	Entertainment News Accuracy	113
5.3	Tourism News Accuracy	114
5.4	Sports News Accuracy	114
5.5	Health News Accuracy	115
5.6	Manual Testing Sore	116
5.7	Overall Accuracy without Text Classifier	118
5.8	Overall Accuracy with Text Classifier	119
5.9	Per domain BLEU Score	120
5.10	Per domain accuracy	120

LIST OF TABLES

Table No.	Table Title	Page No.
1.1	Example of parallel text	20
1.2	Reusable Units	20
2.1	Urdu Alphabets	33
2.2	Urdu digraphs	33
2.3	Punjabi Alphabets	35
2.4	Preposition Grammar Rules in Urdu and Punjabi	36
2.5	Negation in Urdu and Punjabi	37
2.6	Questions in Urdu and Punjabi	38
2.7	Adverb in Urdu and Punjabi	38
2.8	Pronouns in Urdu and Punjabi	39
2.9	Adjectives in Urdu and Punjabi	40
2.10	Nouns in Urdu and Punjabi	41
2.11	Plural in Urdu and Punjabi	42
2.12	Gender in Urdu and Punjabi	42
2.13	Numbers in Urdu and Punjabi	43
2.14	Tense compression of Urdu and Punjabi Sentences	44
3.1	Machine Translation Systems for Indian Languages	70
4.1	Space omission examples	88
4.2	Space insertion examples	81
4.3	Connector alphabets	83
4.4	Non-Connector alphabets	83
4.5	Shape of a connector letter based on different positions	84
4.6	Shape of a Non-connector letter based on different positions	84
4.7	Possible translation	91
4.8	Possible translation with probability values	92
4.9	Izafats used as Proper Nouns	99

Table No.	Table Title	Page No.
4.10	Izafats having 'zer' and 'vao' sounds	100
4.11	Inflections in Urdu	102
4.12	Urdu-Punjabi Character mapping	107
4.13	Urdu-Punjabi Diacritical marks	107
4.14	Urdu words without diacritical marks	108
4.15	Urdu-Punjabi Suffix Mapping	109
5.1	Manually evaluation scores	115
5.2	Testing Data	116
5.3	Confusion matrix of Text classifier	117
5.4	Per Class Recall and Precision	117
5.5	Output of Urdu to Punjabi Translation System	121