

FACTOR-BASED EVALUATION FOR ENGLISH TO HINDI MACHINE TRANSLATIONS

RENU BALYAN



DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2016

©Indian Institute of Technology Delhi (IITD), New Delhi, 2016

FACTOR-BASED EVALUATION FOR ENGLISH TO HINDI MACHINE TRANSLATIONS

by

RENU BALYAN

DEPARTMENT OF MATHEMATICS

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2016

Certificate

It is hereby certified that the thesis entitled “**Factor-based Evaluation for English to Hindi Machine Translations**” submitted by **Ms. Renu Balyan** to the **Indian Institute of Technology Delhi**, for the award of the degree of **Doctor of Philosophy**, is a record of the original *bona fide* research work carried out by her under my supervision and guidance. The thesis work, in my opinion, has reached the requisite standard fulfilling the requirements for the degree of Doctor of Philosophy. The results contained in this thesis have not been submitted in part or in full, to any other university or institute for the award of any degree or diploma.

Dr. Niladri Chatterjee

Professor

Department of Mathematics

Indian Institute of Technology Delhi

Date:

New Delhi

Acknowledgments

This work would not have been in its current shape without the support of many people. It is only because of their confidence in me that I have been able to reach where I am today. I feel this is the best time to express my gratitude to each one of them and let them know how important they all have been during all these years.

First of all, my heartfelt gratitude goes to my guide, supervisor Prof. Niladri Chatterjee for his devotion and guidance, and inspiring me throughout this research. His enormous support not only helped me shape the work reported in this thesis, but also prepared me for my future career. He taught me the basics of writing a research paper and influenced various aspects of my writing and presentation always for the better. Thanks for displaying immense patience while going through the drafts of the papers and the thesis chapters again and again and pin-pointing my mistakes, which probably I never thought to be mistakes.

I was fortunate to visit CNGL, Dublin City University, Ireland as an intern. My special thanks to my supervisors Sudip Kumar Naskar and Antonio Toral for their valuable inputs on my work, while I worked at CNGL. Although it was a short span I learnt a lot from them. I would also like to thank Prof. Josef Van Genabith, CNGL, for providing the required support that helped me attend conferences, so that I could present my research work. This work has been funded in part by the European Commission through the CoSyne project (FP7-ICT-4-248531) and Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the CNGL at DCU, so I convey my thanks to them too for funding part of this research.

I also thank IIT Delhi for providing me with the facilities to carry out my research work. A special mention goes to my SRC committee members for providing valuable comments and suggestions at various stages of my research work. I am also grateful to the anonymous

reviewers for providing valuable feedback and inputs for my research papers. A special thanks to the reviewers of this work for helping me improve the work in a number of ways.

My special thanks to Prof. Vineet Chaitanya and Prof. Deepti Misra (IIIT-Hyderabad), Dr. Sasi Kumar, CDAC, Siva Reddy and Bharat Ambati for providing the linguistic tools and machine translations, that otherwise took too long when translated online. I will be failing my duty if I do not thank some of my colleagues at CDAC, Noida and some students at CDAC and IIT who evaluated the translations that have been part of a number of my experiments conducted during this research.

Last but not the least my family and friends need a special mention too for their sincere support and patience, in particular my parents and in-laws, for being there for my son and husband in my absence. I need to thank my son, Aryan and my husband, Sandeep for their unconditional love, for accepting additional burdens while I was away completing this work, and for supporting me in all my endeavors. They motivated me to finish my work at the earliest possible though things took their own time.

Place: New Delhi

Date:

Renu Balyan

Abstract

In this thesis we address the issues related to automatic evaluation of English to Hindi translations. Although both the languages originated from the Indo-European family, both have undergone many transformations over time due to regional/sub-regional influence. The goal of this work is to develop an evaluation metric that is both meaningful and practically useful. In order to achieve these goals, we follow the diagnostic approach that we term as “*factor-based evaluation*” keeping in line with the latest MT terminology.

Machine translation evaluation (MTE) aims at evaluating translations obtained from machine translation (MT) systems by generating a score for these translations. Although a lot of work has been carried out in this direction for several decades, the MT research community is still in the need of a globally acceptable metric. The existing metrics have faced a lot of criticism as they neither consider words' relevance, nor do they provide any insights into error analysis. Often these may be biased towards a particular MT strategy, and cannot provide any interpretation for the scores generated by them. Moreover, performance of existing state-of-the-art evaluation metrics varies with the language pair under consideration. This observation is more pertinent with respect to translations involving languages of the Indian subcontinent, due to significant differences between the source and the target languages.

As a first step we study the differences between these languages at the grammar level that are important for MT. The next step in this direction is to find the applicability of available diagnostic evaluation tools (DELiC4MT and rgbF) for English to Hindi translations. We propose the addition of phrase-level checkpoints, such as noun compounds (NCs), verb particle constructions (VPCs), and named entities (NEs) on top of the existing PoS-based checkpoints for evaluation of a translation. Further, we work to determine the most common errors produced by the current state-of-the-art online MT systems for English to Hindi using

an automatic tool, *viz.* Herson, along with manual examination. This study helps us establish a preliminary set of error taxonomies for English to Hindi translation outputs. The error taxonomy is categorized as *word level*, *phrase level*, and *sentence level*, apart from *preprocessing errors*. The final evaluation score is computed by analyzing the preprocessing errors and the errors at word level and phrase level.

In this work we also propose a novel set of linguistic and data-driven features for improving and evaluating translations of phrase-based constructs, such as NCs and VPCs. We propose a rule-based approach for determining the translation patterns for 2/3/4-word NCs using *semantic relations* between the nouns of a noun compound. For the VPCs we formulate rules for identifying them and generating their translation patterns based on the *semantic categories* of the subject(s) and/or objects(s) in the sentence.

We finally propose an aggregation scoring formula to integrate scores from linguistic features with penalty scores from human assessments. The penalty scores are computed by using linear regression models built with the help of single classifier, and also an ensemble of classifiers. The models consider all the errors in a translation and assign a weight or penalty to each error. The score generated for a translation is computed as weighted sum of each error score.

To date manual evaluation is considered as the *de facto* standard for MT evaluation. Upon comparing our results with human evaluations we found that the scores computed by the proposed models correlated very well, and these models were capable of simulating the human behavior to a significant extent. Although developed for English to Hindi, we feel this statistics-based scheme will pave the way for development of metrics for other language pairs too.

Table of Contents

Certificate.....	i
Acknowledgments.....	ii
Abstract.....	iv
Table of Contents.....	vi
List of Figures.....	x
List of Tables.....	xii
List of Algorithms.....	xv
List of Equations.....	xvi
1 Introduction	1
1.1 The Motivation.....	1
1.1.1 Differences between English and Hindi (English vs. Hindi).....	3
1.1.2 Issues Related to English to Hindi Translation	8
1.2 Contributions of the Thesis.....	21
1.3 Thesis Outline.....	22
2 Machine Translation and its Evaluation	30
2.1 Machine Translation.....	30
2.1.1 Need for Machine Translation.....	30
2.1.2 MT Systems Classification	32
2.1.3 MT Scenario in India.....	36
2.2 Machine Translation Evaluation.....	39
2.3 Role of Evaluation in MT Context.....	42
2.4 Major Issues in MT Evaluation.....	43
2.5 Essential Criteria for Evaluation Metrics.....	46
2.6 Our Approach: An Experimental Check.....	46
2.7 Chapter Summary.....	49
3 MT Evaluation: Literature Survey	51
3.1 The Evolution of Evaluation Schemes.....	51
3.2 Manual vs. Automatic Evaluation.....	52
3.3 Manual Evaluation Approaches.....	53
3.4 Automatic Evaluation Approaches.....	61

3.4.1	Lexical Matching-based Metrics.....	62
3.4.2	Syntax- and Semantic-based Metrics.....	74
3.4.3	Machine Learning-based Metrics.....	86
3.5	Chapter Summary.....	89
4	Diagnostic Evaluation	91
4.1	Introduction.....	92
4.2	DELiC4EHMT: Diagnostic Evaluation for English to Hindi.....	94
4.2.1	Linguistic Checkpoints.....	95
4.2.2	The Architecture.....	101
4.2.3	PoS Tagging, Text Analysis and Conversion into KAF.....	102
4.2.4	Kybot Profiles for Linguistic Checkpoints.....	107
4.2.5	Evaluation using Linguistic Checkpoints.....	112
4.3	Evaluation using DELiC4EHMT.....	113
4.3.1	Experimental Setup.....	113
4.3.2	Results & Discussions.....	113
4.3.3	Comparison with the state-of-the-art Evaluation Metrics.....	117
4.4	Drawbacks of the Approach.....	118
4.5	rgbF for English to Hindi Translation Diagnostic Evaluation.....	122
4.6	Chapter Summary.....	125
5	MT Error Identification and Classification	127
5.1	Error Identification.....	128
5.1.1	Error Taxonomies and MT Error Typologies.....	129
5.1.2	English to Hindi Translation Errors.....	131
5.1.3	Manual Analysis of E-H Translation Errors.....	139
5.1.4	Error Ranking.....	145
5.2	Error Classification.....	148
5.2.1	Word-Level Errors.....	149
5.2.2	Phrase-Level Errors.....	155
5.2.3	Sentence-Level Errors.....	157
5.2.4	Preprocessing Errors.....	159
5.3	Chapter Summary.....	160
6	Analysis of Phrase-Level Errors	162
6.1	Noun Compounds.....	162

6.1.1	Noun Compound Translation Issues.....	164
6.1.2	Noun Compounds and Semantic Relations.....	167
6.1.3	Translation Patterns for 2-word NCs using Semantic Relations.....	170
6.1.4	Bracketing and Translation Patterns for 3-word and 4-word NCs.....	177
6.1.5	Experimental Results.....	184
6.1.6	Noun Compounds Summary.....	193
6.2	Verb Particle Constructions (VPCs).....	194
6.2.1	Translation Issues in VPCs.....	197
6.2.2	VPCs Identification Issues.....	200
6.2.3	English VPCs Identification.....	202
6.2.4	English VPCs Translation into Hindi.....	206
6.2.5	Algorithmic Implementation & Experimental Results.....	212
6.2.6	VPCs Summary.....	214
6.3	Chapter Summary.....	215
7	Modeling MT Errors	217
7.1	Introduction.....	217
7.2	The Strategy.....	220
7.2.1	Probable Sentence Errors.....	221
7.2.2	Determining Error Weights/ Penalty.....	222
7.3	Regression for Weight/ Penalty Determination.....	231
7.3.1	Experiments: Model Building.....	231
7.3.2	Single Error Models.....	235
7.3.3	Multiple Errors Models.....	236
7.3.4	Combined Errors Models.....	239
7.3.5	Models for Errors Extracted Automatically.....	241
7.4	RMSEs Comparison for Classifiers.....	248
7.5	Comparison with the State-of-the-Art Evaluation Metrics.....	250
7.6	Chapter Summary.....	252
8	Conclusions & Future Work	255
8.1	Summary and Findings of this Thesis.....	255
8.2	Directions for Further Work.....	269
	Bibliography.....	272
	Appendices.....	298

List of Publications.....	308
Brief CV.....	309

List of Figures

2.1	Human and Machine Translation based on Human Degree Interaction.....	33
2.2	The Vauquois Triangle.....	34
2.3	Evaluation Methods in MT Development Cycle.....	44
4.1	Architecture for DELiC4EHMT.....	102
4.2	Sample KAF file for an English Sentence.....	104
4.3	Sample KAF file for a Hindi Sentence.....	106
4.4	Kybot Profile for combined verb particle constructions.....	108
4.5	Kybot Profile for 2-word noun compounds.....	109
4.6	Kybot Profile for a noun phrase (DT JJ* NN*).....	109
4.7	Kybot output file for verb particle construction linguistic checkpoint.....	110
4.8	Sample output for a target match for the VPC “miss out”.....	111
4.9	The 4-gram F-Scores for all the units of different MT systems.....	124
4.10	The trigram F-Scores for all the units of different MT systems.....	125
5.1	Workflow of Automatic Error Classification by Hjerson.....	134
5.2	Error Counts for all the Error Categories for different MT Systems.....	136
5.3	Number of Sentence with length 0-10 containing errors.....	141
5.4	Number of Sentences with length ≤ 10 and > 10 containing errors.....	142
5.5	Aggregate Error Percentages for all the MT systems.....	143
5.6	Hierarchical Error Classification for E-H translations.....	149
6.1	Scheme for handling Translation of Noun Compounds.....	168
6.2	Paraphrase Generation.....	172
6.3	Hierarchy of Partial Semantic Categories.....	211
7.1	A Classifier Ensemble of Regression Models.....	232
7.2	RMSE vs. NumIterations for Bagging: Linear Regression & Bagging: SMOreg.....	233
7.3	The % of Top 22 model-wise ‘good instances’.....	240
7.4	The performance of models for different thresholds for automatically extracted errors.....	248
C.5.1	Top Ten Errors in the MT Systems that Contribute the Most.....	300
C.5.2	Bottom Five Errors in the 4 MT Systems that Contribute the Least	300
C.5.3	Error percentages for MT system 1 (Rule-based System).....	301
C.5.4	Error percentages for MT system 2 (Statistical System).....	301
C.5.5	Error percentages for MT system 3 (Statistical System).....	302

C.5.6	Error percentages for MT system 4 (Hybrid System).....	302
D.6.1	Berkley parser output.....	304
D.6.2	CMU link parser output.....	304
D.6.3	Enju parser output.....	304
E.7.1	RMSEs for Single Errors.....	306
E.7.2	RMSEs for Multiple Errors.....	306
E.7.3	RMSEs for Combined Errors.....	307

List of Tables

1.1	Automatic and Manual Evaluation Scores for Sample Sentences.....	2
3.1	ARPA Adequacy and Fluency scale.....	57
3.2	Meaning Maintenance from Eck and Hori (2005).....	59
3.3	Interpretation of Clarity scores.....	60
4.1	Linguistic Checkpoints for English to Hindi Translation.....	100
4.2	DELiC4EHMT scores for word-level checkpoints for different MT systems.....	114
4.3	DELiC4EHMT scores for phrase-level checkpoints for different MT systems.....	115
4.4	DELiC4EHMT scores for NE Checkpoint.....	116
4.5	Summary of DELiC4EHMT word, phrase, system-level and NE scores.....	117
4.6	Automatic Evaluation Metric scores for MT systems.....	117
4.7	Pearson correlation coefficients between DELiC4EHMT scores and automatic evaluation metrics.....	118
4.8	rgbF unit-wise F-scores for 4-grams and trigrams.....	123
4.9	rgbF overall Precision, Recall and F-scores for 4-grams and trigrams for MT systems.....	124
5.1	Error categories for English to French translation.....	128
5.2	Example of a Hindi Reference Sentence, its Base form and PoS tags.....	135
5.3	Hjerson Raw Error Counts and Error Rates for E-H MT Output.....	136
5.4	Word and Block Error counts and Error Rates.....	138
5.5	Three highest and lowest Error ranking for varying Sentence Lengths.....	147
6.1	Statistics of Noun Compounds in the Corpus.....	169
6.2	Seed Verbs associated with the Semantic Relations.....	171
6.3	Semantic Relations and the Hindi Translation Patterns.....	175
6.4	Adjacency and Dependency Models for Frequency and Probability based Approaches.....	179
6.5	Sample1 NC for which the Adjacency and Dependency models contradict.....	180
6.6	Sample2 NC for which the Adjacency and Dependency models contradict.....	180
6.7	3-word NC Bracketing Accuracy.....	186
6.8	Percentages of errors occurring in the NCs translated by the translators.....	189
6.9	Translation outputs of various translators for some example noun compounds in a	

	sentence.....	191
6.10	Precision, Recall and F-Score for the different translators.....	193
6.11	Number of VPCs with frequency in BNC Corpus.....	195
6.12	VPCs translations by the translators.....	197
6.13	Different senses and translations of the VPC “put up”.....	199
6.14	VPCs not identified by the Stanford parser.....	201
6.15	VPC Identification statistics.....	202
6.16	VPC Identification Rules and Example Sentences.....	204
6.17	VPC identification for different sizes of unseen and seen dataset.....	206
6.18	Number of VPCs with WordNet Senses.....	207
6.19	WordNet sense and Hindi Verbs for “break down”.....	208
6.20	# of WordNet (WN) sense and # of Hindi verbs.....	208
6.21	Accuracy, Precision, Recall and F-Score for different translators.....	213
7.1	The List of Errors for English to Hindi Translation.....	221
7.2	Descriptive Statistics for Sentences having words >10.....	223
7.3	Descriptive Statistics for Sentences having words <=10.....	224
7.4	Two-way ANOVA results for Errors vs. Sentences.....	228
7.5	The Estimated Marginal Means for Errors.....	229
7.6	Codes used for Model Generation.....	234
7.7	# of sentences for Single Errors that are within Threshold.....	236
7.8	# of sentences for Multiple Errors that are within Threshold.....	237
7.9	Model and Error weights for top five models for multiple errors.....	238
7.10	# of sentences for Combined Errors that are within Threshold.....	240
7.11	Errors Extracted Automatically using Existing Linguistic Tools.....	241
7.12	Performance of models for Single Errors Extracted Automatically	243
7.13	Performance of models for Multiple Errors Extracted Automatically	244
7.14	Model and Error weights for top five models for multiple errors.....	245
7.15	Performance of models for Combined Errors Extracted Automatically	246
7.16	Model and Error weights for top five models for Combined Errors.....	247
7.17	RMSEs for all the classifiers.....	249
7.18	Two-way ANOVA results for the RMSEs.....	250
7.19	Pearson Correlation for state-of-the-art metrics and our Models for manually identified Errors.....	251

7.20	Pearson Correlation for state-of-the-art metrics and our Models for automatically extracted Errors.....	252
A.1.1	Summary of the major MT projects in India.....	298
B.4.1	Reference file example of Hindi sentence in rgbF format.....	299
D.6.1	Semantic labels proposed in the literature.....	303
D.6.2	Precision, Recall and F-Score for different parsers.....	305
E.7.1	The Estimated Marginal Means for Sentences.....	307

List of Algorithms

6.1	Semantic Relation Identification for a 2-word English Noun Compound.....	173
6.2	Translation Pattern Generation for a 2-word English Noun Compound.....	176
6.3	Translation Pattern Generation for a 3-word English Noun Compound.....	181
6.4	VPC Identification and Translation.....	212

List of Equations

2.1	Linear Model after Aggregation.....	48
2.2	Linear Model of all Individuals.....	48
3.1	Translation Error/Edit rate.....	64
3.2	BLEU score.....	65
3.3	BLEU Brevity Penalty.....	65
3.4	BLEU Modified n-gram precision.....	65
3.5	NIST score.....	68
3.6	NIST Info weight.....	68
3.7	Precision and Recall for GTM.....	69
3.8	F-score for GTM.....	70
3.9	ROUGE-N score.....	70
3.10	Precision, Recall and ROUGE-L score.....	71
3.11	Precision, Recall and ROUGE-S score.....	71
3.12	METEOR score.....	72
3.13	METEOR Fmean.....	72
3.14	METEOR Penalty.....	72
3.15	Syntactic Tree Matching.....	75
3.16	Head-Word Chain Matching.....	76
3.17	BLEUÂTRE Recall score.....	76
3.18	BLEUÂTRE Dependent ordering score.....	76
3.19	BLEUÂTRE Length Penalty.....	77
3.20	LEPOR score.....	80
3.21	Length and nPoS Penalty for LEPOR.....	80
3.22	N-gram Position Difference Penalty for LEPOR.....	80
3.23	Extended Length Penalty.....	81
3.24	hLEPOR score.....	81
3.25	MaxSim Similarity Score.....	82
3.26	QUEEN score.....	88
3.27	ULC score.....	89
4.1	Recall for Checkpoints.....	112