FACTOR-BASED EVALUATION FOR ENGLISH TO HINDI MACHINE TRANSLATIONS

RENU BALYAN

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 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 (DELiC4MTand 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*. Hjerson, 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

List of Figures

List of Tables

List of Algorithms

List of Equations