CONTRIBUTIONS TO ENGLISH TO HINDI MACHINE TRANSLATION USING EXAMPLE-BASED APPROACH

by

DEEPA GUPTA

Department of Mathematics

Submitted

in fulfilment of the requirement of

the degree of

Doctor of Philosophy

to the

Indian Institute of Technology Delhi

Hauz Khas, New Delhi-110016, India

January, 2005

Dedicated to

My Parents,

My Brother Ashish and

My Thesis Supervisor...

Certificate

This is to certify that the thesis entitled "**Contributions to English to Hindi Machine Translation Using Example-Based Approach**" submitted by **Ms. Deepa Gupta** to the Department of Mathematics, Indian Institute of Technology Delhi, for the award of the degree of Doctor of Philosophy, is a record of *bona fide* research work carried out by her under my guidance and supervision.

The thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The work contained in this thesis has not been submitted to any other university or institute for the award of any degree or diploma.

Dr. Niladri Chatterjee
Assistant Professor
Department of Mathematics
Indian Institute of Technology Delhi
Delhi (INDIA)

Acknowledgement

If I say that this is my thesis it would be totally untrue. It is like a dream come true. There are people in this world, some of them so wonderful, who helped in making this dream, a product that you are holding in your hand. I would like to thank all of them, and in particular:

Dr. Niladri Chatterjee - mentor, guru and friend, taught me the basics of research and stayed with me right till the end. His efforts, comments, advices and ideas developed my thinking, and improved my way of presentation. Without his constant encouragement, keen interest, inspiring criticism and invaluable guidance, I would not have accomplished my work. I admit that his efforts need much more acknowledgement than expressed here.

I acknowledge and thank the Indian Institute of Technology Delhi and Tata Infotech Research Lab who funded this research. I sincerely thank all the faculty members of Department of Mathematics, especially, I express my gratitude for Prof B. Chandra and Dr. R. K. Sharma, for providing me continuous moral support and help. I thank my SRC members, Prof. Saroj Kaushik and Prof. B. R. Handa, for their time and efforts. I also thank the department administrative staff for their assistance. I extend my thanks to Prof. R. B. Nair and Dr. Wagish Shukla of IIT Delhi, and Prof. Vaishna Narang, Prof. P. K. Pandey, Prof. G. V. Singh, Dr. D. K. Lobiyal. and Dr. Girish Nath Jha of Jawaharlal Nehru University Delhi, for the enlightening discussions on basics of languages.

I would like to express my sincere thanks to my friends Priya and Dharmendra for many fruitful discussions regarding my research problem. I thank Mr. Gaurav

Kashyap for helping me in the implementation of the algorithms. In particular, I would like to thank Inderdeep Singh, for his help in writing some part of the thesis. I want to give special thanks to my friends, Sonia, Pranita and Nutan, for helping me in both good and bad times. I would like to thank Prabhakhar for his brotherly support. I extend my thanks to Manju, Anita, Sarita, Subhashini and Anju for cheering me, always.

Shailly and Geeta - amazing friends who read the manuscript and gave honest comments. Both of them also stayed with me in the process, and handled me, and sometimes my out-of-control emotions so well. Especially, I wish to extend my thanks to Geeta for providing me stay in her hostel room, and also for her wonderful help when my leg got fractured when we knew each other for a month only. I wish to acknowledge Krishna for his constant help, both academic and nonacademic, and his continuous encouragement.

I convey my sincere regards to my parents, and brothers for the sacrifices they have made, for the patience they have shown, and for the love and blessing they have showered. I thank Arun for his moral support. Most imperative of all, I would like to express my profound sense of gratitude and appreciation to my sister Neetu. Her irrational and unbreakable belief in me bordered on craziness at times.

I cannot avoid to mention my friend Sharad who deserves more than a little acknowledgement. His constant inspiration and untiring support has sustained my confidence throughout this work.

Finally, I thank GOD for every thing.

Deepa Gupta

Abstract

This research focuses on development of Example Based Machine Translation (EBMT) system for English to Hindi. Development of a machine translation (MT) system typically demands a large volume of computational resources. For example, rule-based MT systems require extraction of syntactic and semantic knowledge in the form of rules, statistics-based MT systems require huge parallel corpus containing sentences in the source languages and their translations in target language. Requirement of such computational resources is much less in respect of EMBT. This makes development of EBMT systems for English to Hindi translation feasible, where availability of large-scale computational resources is still scarce. The primary motivation for this work comes because of the following:

a) Although a small number of English to Hindi MT systems are already available, the outputs produced by them are not of high quality all the time. Through this work we intend to analyze the difficulties that lead to this below par performance, and try to provide some solutions for them.

b) There are several other major languages (e.g., Bengali, Punjabi, Gujrathi) in the Indian subcontinent. Demand for developing MT systems from English to these languages is increasing rapidly. But at the same time, development of computational resources in these languages is still at its infancy. Since many of these languages are similar to Hindi, syntactically as well as lexicon wise, the research carried out here should help developing MT systems from English to these languages as well.

The major contributions of this research may be described as follows:

1) Development of a systematic adaptation scheme. We proposed an adaptation scheme consisting of ten basic operations. These operations work not only at word level, but at suffix level as well. This makes adaptation less expensive in many situations.

2) Study of Divergence. We observe that occurrence of divergence causes major difficulty for any MT systems. In this work we make an in depth study of the different types of divergence, and categorize them.

3) Development of Retrieval scheme. We propose a novel approach for measuring similarity between sentences. We suggest that retrieval strategy, with respect to an EBMT system, will be most efficient if it measures similarity on the basis of cost of adaptation. In this work we provide a complete framework for an efficient retrieval scheme on the basis of our studies on "divergence" and "cost of adaptation".

4) Dealing with Complex sentences. Handling complex sentences by an MT system is generally considered to be difficult. In this work we propose a "split and translate" technique for translating complex sentences under an EBMT framework.

We feel that the overall scheme proposed in this research will pave the way for developing an efficient EBMT system for translating from English to Hindi. We hope that this research will also help development of MT systems from English to other languages of the Indian subcontinent.

Contents