Universal Networking Language Based Analysis and Generation for Bengali Case Structure Constructs

Kuntal Dey[1] and Pushpak Bhattacharyya[2]

[1] Veritas Software, Pune, India.
u2ckuntal@yahoo.com
[2] Computer Science and Engineering Department
Indian Institute of Technology, Bombay, India.
pb@cse.iitb.ac.in

Abstract. Case structure analysis forms the foundation for any natural language processing task. In this paper we present the computational analysis of the complex case structure of Bengali- a member of the Indo Aryan family of languages- with a view toward interlingua based MT. Bengali is ranked 4^{th} in the list of languages ordered according to the size of the population that speaks the language. Extremely interesting language phenomena involving morphology, case structure, word order and word senses makes the processing of Bengali a worthwhile and challenging proposition. A recently proposed scheme called the *Universal Networking Language* has been used as the interlingua. The approach is adaptable to other members of the vast Indo Aryan language family. The parallel development of both the analyzer and the generator system leads to an insightful intra-system verification process in place. Our approach is *rule based* and makes use of authoritative treatises on Bengali grammar.

1 Introduction

Bengali is spoken by about 189 million people and is ranked 4^{th} in the world in terms of the number of people speaking the language (ref: *http://www.harpercollege.edu/~mhealy/g101ilec/intro/clt/cltclt/top100.html*). Like most languages in the Indo Aryan family, descended from Sanskrit, Bengali has the SOV structure with some typical characteristics. A motivating factor for creating a system for processing Bengali is the possibility of laying the framework for processing many other Indian languages too.

Work on Indian language processing abounds. *Project Anubaad* [1] for machine translation from English to Bengali in the newspaper domain uses the *direct translation approach*. *Angalabharati* [2] system for English Hindi machine translation is based on pattern directed rules for English, which generates a *pseudo-target-language* applicable to a group of Indian Languages. In MATRA [3], a web based MT system for English to Hindi in the newspaper domain, the input text is transformed into case-frame like structures and the the target

Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 215–229.