Chapter 3
UNL-Based Bangla Machine Translation Framework

Nawab Yousuf Ali
East West University, Bangladesh

Shamim H. Ripon
East West University, Bangladesh

ABSTRACT

The usage of native language through Internet is highly demanding due to the rapid increase of Internet-based applications in daily life. As information is available in the Internet in different languages, it is impossible to retrieve the information in other languages. Universal Networking Language (UNL) addresses this issue by converting the requested information from other languages to UNL expressions followed by UNL expressions to respective native languages. Even though Bangla is the sixth most popular language in the world, there is no system developed so far to convert Bangla text into UNL expressions and vice versa. For this purpose, the authors develop a framework. The framework has two constituent parts: 1) EnConverter: converts Bangla native sentences into UNL expressions considering UNL compatible Bangla word dictionary and analysis rules, and 2) DeConverter: converts UNL expressions into respective Bangla sentences considering Bangla word dictionary and generations rules. In both cases, case structure analysis, Bangla parts of speech, and different forms of verbs along with their prefixes, suffixes, and inflections are taken into consideration. This chapter describes the complete theoretical analyses of the EnConversion and DeConversion frameworks. The experimental results confirm that the proposed framework can successfully convert Bangla sentences into UNL expressions, and also can convert UNL expressions into corresponding Bangla sentences.

DOI: 10.4018/978-1-4666-3970-6.ch003

1. INTRODUCTION

Universal Networking Language (UNL) is a project under the auspices of the United Nations University (UNU), Tokyo, Japan. The mission of the UNL project is to allow people across nations to access information in the Internet in their own languages (Uchida, Zhu & Senta, 2005). Hundreds of millions of people of almost all levels of education and attitudes of different jobs all over the world use the Internet for different purposes (Ali, Das, Mamun & Nurannabi, 2008). The last decade of the 20th century witnessed an unimaginary acceleration in the development of information technology in all fields of life. The decade also witnessed a great increase in the spread and popularity of the Internet. English is the main language of the Internet. Understandably not all people know English. Teeming millions are deprived to access the information repositories directly in native language. On the other hand, vast information resources in different languages could not be shared. Knowledge and information are scattered all over the world and remain mostly inaccessible due to non-machine representation and language barrier (Ali, Das, Mamun & Choudhury, 2008). Translation is the only means to disseminate information but only with much effort and involving direct and indirect cost. Language barrier hinders progress at individual level, institutionally and nationally although nations are becoming more interdependent and need to exchange information. Knowledge sources are to be shared globally as much as possible to advance civilization.

Among those who did their best to tackle this problem was the United Nations University/Institute of Advanced Studies (UNU/IAS). The institute conducted a review of all internationally available machine translation programs and finally decided to start devising a better, more efficient and more workable technique to develop a human language neutral meta-language for Internet. The result of the project is Universal Networking Language (UNL) (Chudhury, Ali, Sarkar & Ahsan, 2005). The UNL project is a large scale international cooperation with the goal to provide information in the Internet in all national languages of the members of the United Nations. The goal is to eliminate the massive task of translation between two languages and reduce language to language translation to a one time conversion to UNL. Once information written in one language is "enconverted" into UNL it will be able to be shared by anyone in the world (Ali, Das, Mamun & Choudhury, 2008). That means the UNL is based on developing an intermediary language system whereby any written text can be converted to many languages and simultaneously, all texts written in different languages can be converted to that particular language. For example, Bangla corpora, once converted to UNL, can be translated to any other language given UNL system built for that language shown in Figure 1.

The UNL system does this by representing only the semantics of a native language sentence in a hypergraph. The hypergraph is composed of nodes connected by semantic relations. Nodes or Universal Words (UWs) are words loaned from English and disambiguated by their positioning in a knowledge base (KB) of conceptual hierarchies. Function words, such as determiners and auxiliaries are represented as attributes to UWs or nodes to provide additional information. English is used for UW, attributes and relations. Enconverter (Enconverter, 2002) converts each native language sentence to a UNL hypergraph and Deconverter (Deconverter, 2002) translates from hypergraph to any native language. The hypergraph has formal English text realization termed as UNL document (like HTML or XML). The development of the language specific components - dictionary and analysis rules - is carried out by researchers across the world.

The difficulty in these translation systems lies in the language analysis process to be performed by the computer in analyzing a sentence in its semantic representation. The computer has to discriminate the lexical and syntactic ambiguities, and