Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

The EB-ANUBAD translator: A hybrid scheme

SAHA Goutam Kumar

(*Centre for Development of Advanced Computing, Kolkata, West Bengal 700091, India*)

E-mail: gksaha@rediffmail.com; sahagk@gmail.com

Received Aug. 8, 2005; revision accepted Sept. 5, 2005

Abstract: This article is aimed at describing a hybrid scheme for English to Bangla translation. The translated output in English scripts is useful for learning Bengali language. This is a significant contribution to Human Language Technology generation also. About two hundred million people in West Bengal and Tripura (two states in India) and in Bangladesh (a country whose people speak and write Bangla as their first language). This proposed translator would benefit Bengalee society because rural people are not usually very conversant with English. The English to Bangla Translator is being enhanced. This system (English-Bangla-ANUBAD or EB-ANUBAD) takes a paragraph of English sentences as input sentences and produces equivalent Bangla sentences. EB-ANUBAD system is comprised of a preprocessor, morphological parser, semantic parser using English word ontology for context disambiguation, an electronic lexicon associated with grammatical information and a discourse processor, and also uses a lexical disambiguation analyzer. This system does not rely on a stochastic approach. Rather, it is based on a special kind of hybrid architecture of transformer and rule-based Natural Language Engineering (NLE) architectures along with various linguistic knowledge components of both English and Bangla.

Key words: Machine translation, Rule based, Transformation based, Natural Language Engineering (NLE) System
doi:10.1631/jzus.2005.A1047 **Document code:** A **CLC number:** TP390

INTRODUCTION

Bangla language is characterized by a rich system of inflections (VIBHAKTI), derivation, and compound formation (Saha *et al*., 2004; Dash, 1994; Chakroborty, 2003), which is why the NLE using Bangla (output generation) is a very challenging task.

Natural Language Engineering (NLE) is the process of computer analysis of input provided in a human language (natural language) and conversion of this input into a useful form of representation. The input of an NLP system can be: written text or speech. This paper deals with the written text only. In order to process written text, we need: (a) lexical, (b) syntactic, (c) semantic knowledge about the language and (d) discourse information along with real world knowledge.

The purpose of lexical processing is to determine meanings of individual words. Syntactic analysis deals with syntactic structure. Semantic analysis deals with the context-independent meaning representation whereas discourse processing deals with final meaning representation.

The term ontology simply denotes a group of "concepts" organized to reflect the relationships between the concepts. A lexicographer has the primitive task of building of ontology. Each word forms a class in which more than one entity can be included. Suppose there are words like biscuits, pizza, cake, etc. All these words can be put under a single category i.e., Food (edible one). This type of categorization can be performed through the "is-a-kind-of" relation. Such information is useful for the purpose of context disambiguation. The EB-ANUBAD system uses such ontological analysis also.

The proposed translator (EB-ANUBAD) uses (i) the grammar for the input or source language, (ii) a source-to-target language dictionary, (iii) a set of source-to-target language rules, and (iv) an exception handler.

HYBRID EB-ANUBAD SYSTEM

This English to Bangla translator is based on a special architecture of rule-based and transformer architecture. Fig.1 shows various processing modules of the new hybrid translator system based on 300 rules (more rules are being developed). It is upgraded with linguistic knowledge architecture also and is provided with morphological parser, semantic parser and ontological analyzer, disambiguation processing, and discourse analyzer (Bharati *et al*., 1991; 2000; Terry, 1972; Jurafsky and Martin, 2000; Ma, 2002). The system was developed using VB 6.0 and MS Access 2000.

To begin with, the lexicon is comprised of 4000 English words only. The EB-ANUBAD translator system's user interface for input and output is shown in Fig.3. This interface shows a paragraph of English (input) sentences in the upper text box and the translated paragraph of Bangla (output) sentences in the lower text box.

PROCESSING BLOCKS IN THE EB-ANUBAD SYSTEM

The parse tree for "Light the light light" is shown in Fig.2a. An English sentence follows subject-verb-object (SVO) formation.

Fig.2b shows the corresponding Bangla parse tree based on subject-object-verb (SOV) structure for the input English sentence "Light the light light."

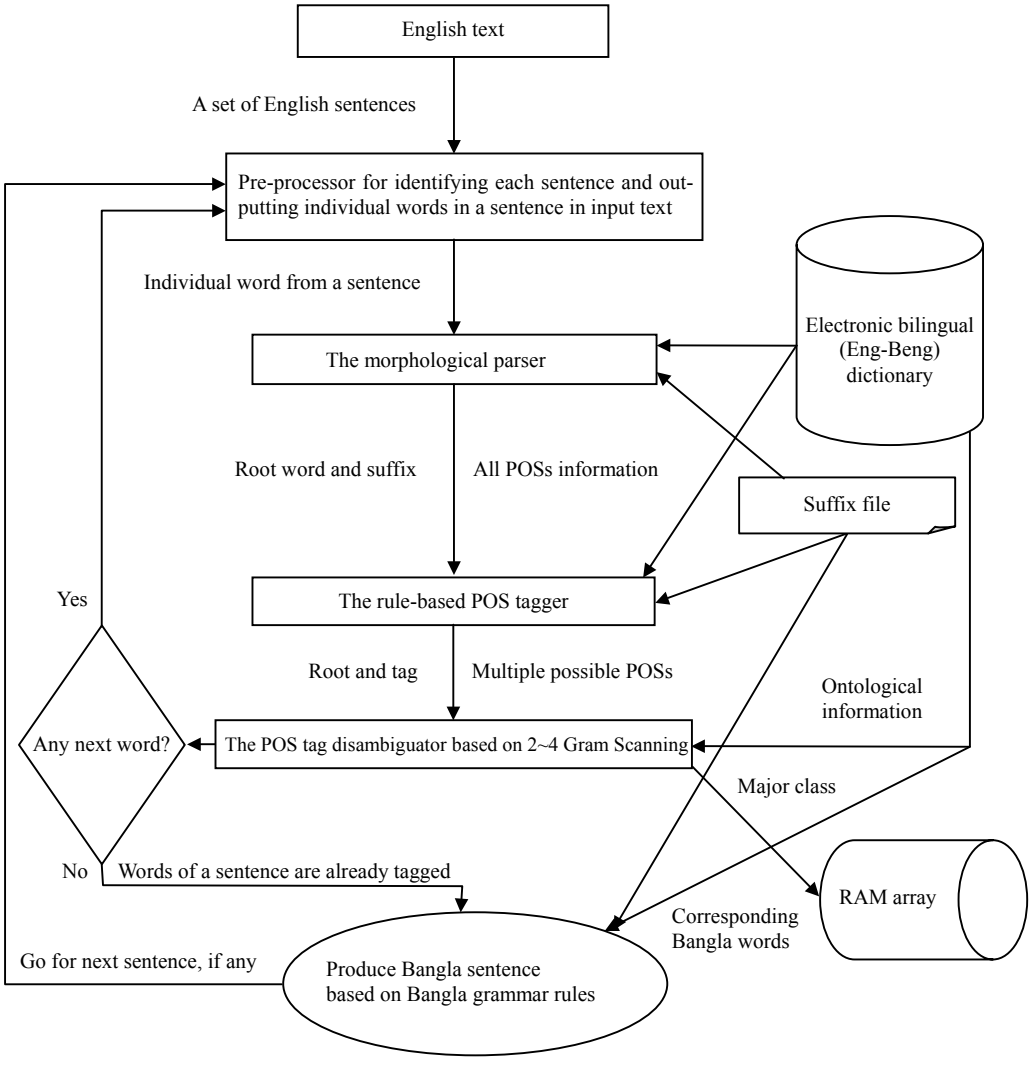

Fig.1 Processing blocks of the EB-ANUBAD system

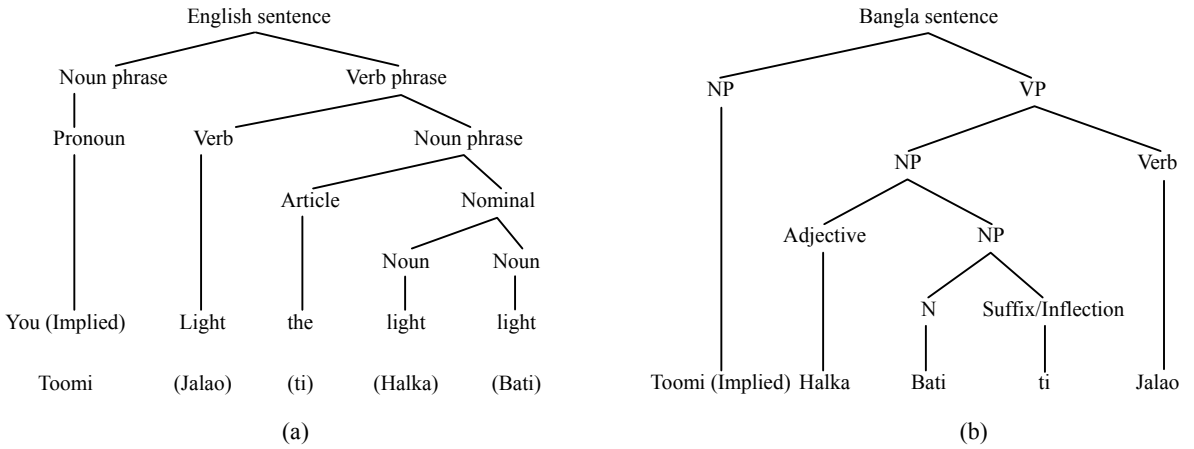

Fig.2 The parse tree (a) and bangle parse tree (b) of "Light the light light"

The English-Bangla bilingual lexicon contains information on corresponding Bangla words, part-of-speech, inflection, ontology, etc. for an English lexicon. The translator's interface is shown in Fig.3. POS ambiguity is resolved through my own rules on using 2~4 Gram bi-directional scanning on an English sentence. Semantic ambiguity is resolved through ontological information and surrounding words. This machine translator is not a transliteration system like the Angla-Bharati system (English to Hindi) in India which simply uses bilingual dictionary to get the target language word for an English word and produces all possible Hindi sentences against an English sentence that has POS and semantic ambiguity. My system is a deterministic system that produces the best one only. It has been tested OK with numerous sentences given by many users.

The system can handle a word that is not present in the lexicon. It can handle lexicon disambiguation (a word with multiple part-of-speech tags or with multiple meanings) also. For example, the word "Light" (in English) has multiple POS tags namely, verb, adjective and noun. Light (verb) is Jalao (in Bangla). Light (Adjective) is Halka (in Bangla). Light (noun) is Baati (in Bangla). Again, the EB-ANUBAD system is capable of context disambiguation also.

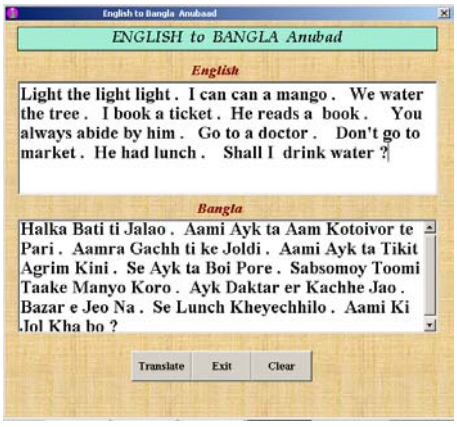

Fig.3 The EB-ANUBAD interface

For example, for the input sentence in English like, "I had a Pizza," the EB-ANUBAD's output is "Aami Ekta Pizza Kheyechhilam," (in Bangla). Or, for the input sentence "I had a dog", the system's output is "Aamar Ekti Kukur Chhilo" (in Bangla). The word "had" has two different context meanings. Again, for an example of POS disambiguation, the system's output is "Aamra Jol Khai" for the input English sentence "We drink water" (water as noun). Or, for the input sentence "Water the tree" (water as a verb), the system's output is "Gaachh Tite Jol Dao". Or, for the input sentence "This is a water tank" (water as an adjective), the system's output is "Eti Joler Tank".

This system does not use any pre-tagged English corpus because it is not a stochastic approach. We have not used Hidden Markov Model (HMM) also. EB-ANUBAD uses its built-in POS tagger only.

CONCLUSION

This system can handle the most challenging "disambiguation" aspects of NLE through semantic net analysis. The EB-ANUBAD translator system is not exactly based on any conventional rule-based,

stochastic or transformation based NLE. This is based on a special kind of hybrid architecture of rule-based and transformer system along with an integrated parser for both morphological and semantic analysis. This system gives only a unique translated output sentence against an English sentence. Much attention is given in developing such a complex NLE translator system to generate a deterministic output sentence for an input or a source sentence. Both the lexicon and context disambiguation processing have been incorporated to work satisfactorily. This system incorporates also various linguistic components like Bangla inflections (Vibhakti), derivation, Karak (endings) and compound formation also. The system is easily upgradable with new grammatical rules and lexicons. Study is going on towards enhancing this translator. It has been tested with various text and we get around 98% correct result. Bangla grammatical inflectional error is found to be about 1%~2%. This is a low cost domain independent translator system aimed at producing reliable output with high performance and relatively high accuracy and dedicated to rural Bengalee people for understanding English text. This hybrid approach (i.e. using a few thousand rules and transformation) can easily be adopted for any source and target languages.

ACKNOWLEDGEMENT

The author (developer) is thankful to Dr. A.B. Saha, Dr. Om Vikas, Dr. S. Ramakrishnan for their encouragement.

References

Bharati, A., Chaitanya, V., Sangal, R., 1991. A computational grammar for Indian languages processing. *Indian Linguistics Journal*, **52**:91-103.

Bharati, A., Chaitanya, V., Sangal, R., 2000. Natural Language Processing. PHI.

Chakroborty, B., 2003. Uchchotoro Bangla Byakaron. Akshay Malancha.

Dash, K.C.(Ed.), 1994. Indian Semantics. Agamakala Publications, Delhi.

Jurafsky, D., Martin, J.H., 2000. Speech and Language Processing. Pearson Education.

Ma, Q., 2002. Natural Language Processing with Neural Networks. Language Engineering Conference, Hyderabad.

Saha, G.K., Saha, A.B., Debnath, S., 2004. Computer Assisted Bangla POS Tagging. iSTRAN, Tata McGraw-Hill, New Delhi.

Terry, W., 1972. Understanding Natural Language. Academic Press, New York.