EtranS-English to Sanskrit Machine Translation

Promila Bahadur
Amity University
Noida (U.P.), India.

A.K.Jain
Indian Institute of Technology
Kanpur (U.P.), India.

D.S.Chauhan
UttrakhandTechnical University
Dehradoon (UK.), India.

ABSTRACT

Machine Translation has been a topic of research from the past many years. Many methods and techniques have been proposed and developed. However, quality of translation has always been a matter of concern. In this paper, we outline a target language generation mechanism in English-Sanskrit Machine Translation using rule based machine translation technique. The methodology for design and development is implemented in the form of software developed "EtranS".

Keywords

Analysis, Machine translation, translation theory, Interlingua, language divergence, Sanskrit, natural language processing.

1. INTRODUCTION

English is a widely spoken language across the global and most official communication and documentation is being done in this language. In India, there exist several regional languages including Hindi, where a lot of documentation exists in this language. The Sanskrit is considered to be mother of all Indian languages and is one of the oldest synthetic language in which a lot of ancient literature exists. Since English is modern day "global language", it has always been a challenge before natural language processing community to find efficient mechanism for this translation pair [2, 3, 4].

We compare and analyze differences between the two languages which are pre-requisite before getting into translation technique. There are four major parameters namely, essence, tense, number and translational equivalence, that are needed to be considered for the translation of this language pair. The essence of English is that it is evolved therefore it is a natural language. Sanskrit is formulated by sages like Panini hence it is an Artificial or Synthetic language. The English language has twelve tenses in all primarily Past, Present and Future. All three have a Perfect, Indefinite, Continuous and Perfect Continuous and it makes twelve forms of tenses. Sanskrit has primarily six tenses, Present, Past, Future, Order, Blessing and Inspiration. The English have two numbers i.e., Singular and Plural whereas Sanskrit has three numbers Singular, Dual and Plural. In general, we can state that the model consists of a set of translation rules to translate from source to target sentence, which is a framework of Rule based Machine Translation System. The rules are framed, keeping in view the grammar of the source and the target language (Translational Equivalence) [2].

2. THE PROCESS ENGINE

The functional approach to translation is developed on the basis of "The Two Way Translation Model"[1] shown in Fig. 2. This model states that for the translation from source to target language first Top Down and then Bottom UP approach is adopted. It presents a simple technique for translation. There are two phases, the first phase follows, the Top Down approach. Here, we begin with syntax analysis, followed by semantic analysis and then mapping of tokens is done, which are generated during syntax analysis. The second phase, does Bottom to Top analysis. It begins with intermediate process of mapping, felicitated by first phase, which is followed by morphological analysis and finally target language is generated.

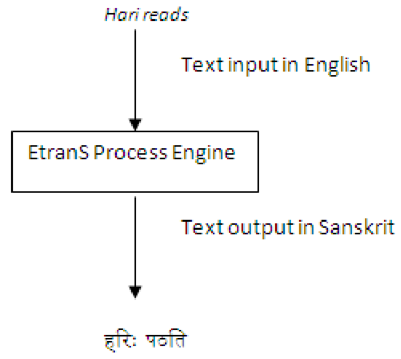


Fig 1: Etrans Process Engine

The EtranS process engine displayed in Fig.1 takes input in source language and generates output in target language, in our case it is English and Sanskrit. The engine has two major components

i. The parsing process
ii. The generator process

3. THE PARSING PROCESS

Parsing process is the first component of the process engine. This component is responsible for the Top to Bottom analysis. It has following sub processes

i. The Input Process
ii. Sentence Analyzer Process
iii. Morphological Analysis Process
iv. The EtranS Lexicon
v. The Parsing Process

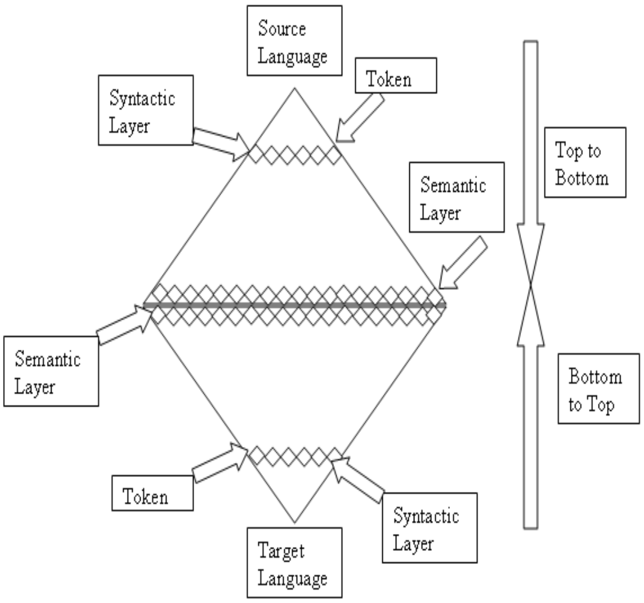


Fig 2:Showing the Translation Model

3.1 The Input Process

This process is the first small step towards translation process. It takes sentence as input in a text box developed in a GUI within the software developed for the translation process.

3.2 Sentence Analyzer Process

This process analyses the sentence and states the category of the input sentence i.e., whether the sentence is small, large or extra large. The sentence is divided into tokens (e.g. Ram goes to the market. For this input sentence tokens, are generated like ram, goes, to, the, market). Tokens can also have more than one word such as, take off, in case of. It also extracts the root words from the tokens e.g., from “goes” “go” is extracted .The tokens are identified and morphological analysis is done, as shown in table.1.

3.3 Morphological Analysis Process

The lexicon or the database developed plays a pivotal role for Morphological analysis. As it searches through the lexicon to gather above mentioned categories of part of speech and its sub categories

This process takes the tokens as input and gathers grammatical information on that e.g., Morphological analysis for tokens like Ram, is, eating would provide following information

3.4 The EtranS Lexicon

A bilingual lexicon is developed as a bridge between the source as well as the target language [1]. Structure of the lexicon contains various categories and sub categories pertaining to the source and the target language, as shown in table 2.

Table 1: Analysis of tokens

Table 2: Major categories of Lexicon tokens

3.5 The Parsing Process

This process checks whether the input sentence is grammatically correct or not, as shown in Fig. 3 and 4 with the help of EtranS rule bank. The information gathered from above mentioned process helps in analyzing the grammatical aspect of the sentence and on the basis of the rules assessment is done for e.g. for sentences “Ram is eating” and ”Ram are eating”, we have the following analysis:

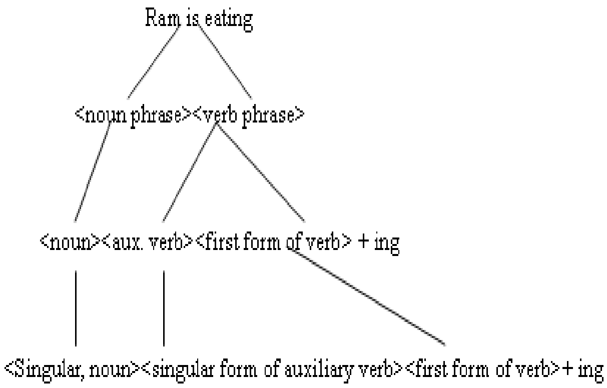

Ram is eating
<noun phrase><verb phrase>
<noun><aux. verb><first form of verb> + ing
<Singular, noun><singular form of auxiliary verb><first form of verb>+ ing

Fig 3: Tree Showing analysis of ‘Ram is going”

i. (Note-This is a rule in Present Continuous tense)
ii. Here, we can see that the sentence stands to be true both at morphological as well as parser level.
iii. Therefore, this sentence is correct.

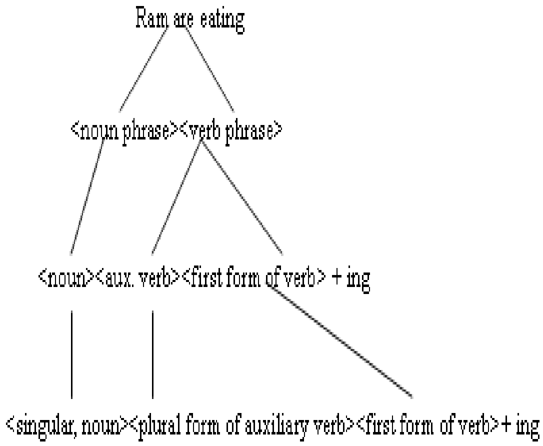

Ram are eating
<noun phrase><verb phrase>
<noun><aux. verb><first form of verb> + ing
<singular, noun><plural form of auxiliary verb><first form of verb>+ ing

Fig 4: Tree Showing analysis of “Ram are going”

i. (Note-This is not a matching rule in Present Continuous tense)
ii. Here, we can see that the sentence stands to be true at morphological level but false at parser level.
iii. Therefore, this sentence is not correct.

Noun	Verb	Root Word
Ram- name of a person Proper noun , Masculine gender, Singular	Is – auxiliary verb Eating-to eat transitive verb, animate, continuous word	Eat

Category	Information stored
Noun	Person, gender ,number, root word
Verb	Tense , gender

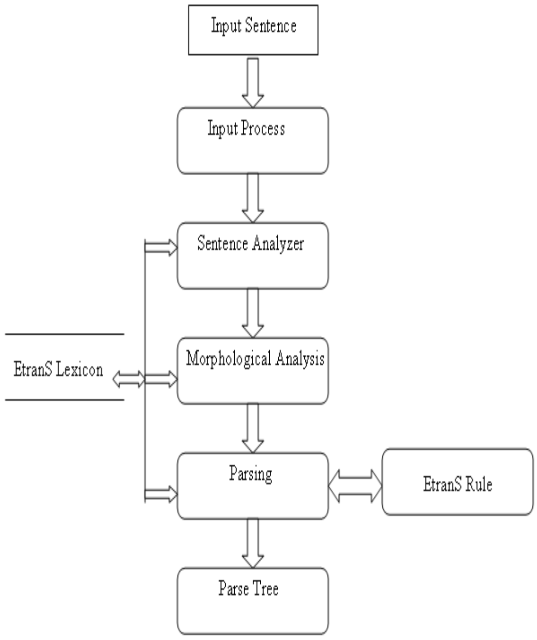


Fig 5: Parsing Process

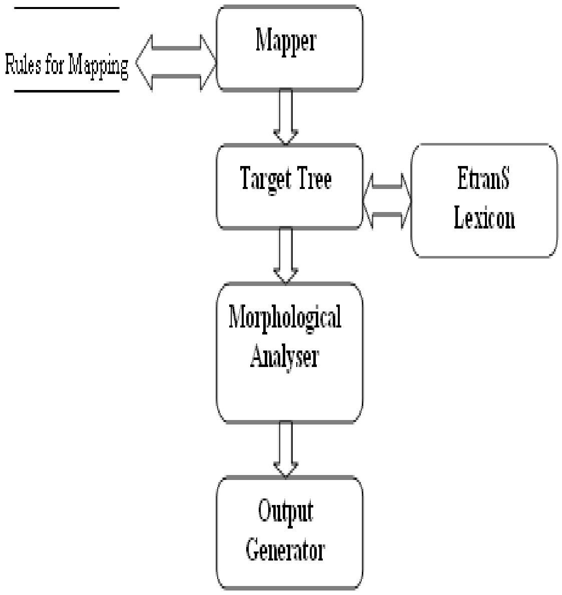


Fig 6: Generator Process

4. The Generator Process

This is a second phase of translation process and does bottom to top analysis. It is composed of following processes

i. Mapping Process
ii. Morphological Analysis
iii. Output Process

4.1 Mapping Process

Mapping is done purely on the basis of the information passed from the morphological module. Since, in Sanskrit a word contains conjunction, preposition and other information therefore this needs to be considered while mapping process, while, English have preposition and conjunction in separate, as shown in Fig. 7 and 8.

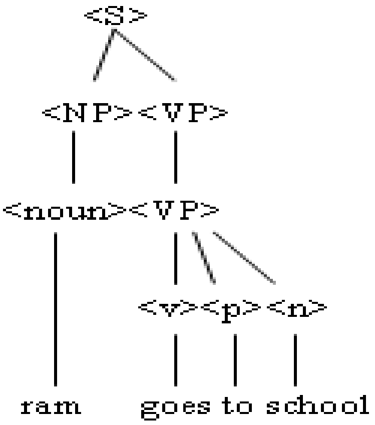


Fig 7:Tree generated in source langauge

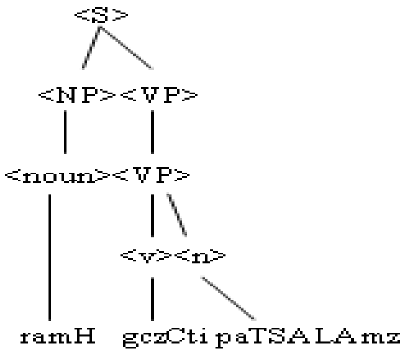


Fig 8:Mapping Tree generated in target langauge*

*Trees showing translation (mapping)process from source to target language

4.2 Morphological Analysis

This module searches the root word in the target language and maps similar information to complete the translation process. For example, "Ram goes to school" would be translated as "*ramH gczCti paTSAMAmz*".Here ram would be searched as root word and "*H"* would be added as part of singular noun and *pratham vibhakti* to form "*ramH*". Further, *paTSAMAmz*[1] would be searched as it is neutral gender and singular in number having *dritiya vibhakti*. and this would be followed by search of goes which have root word go and ram would be added as it is singular -present tense. As we can see in the Fig.7 that ram is a singular noun therefore in Fig. 8 it is mapped with ramH (ram+H) [In Romanized form] which means root word + vibhakti [1].

4.3 Output Process

This process gathers the information from the leaves of the trees generated after above processes. The leaves contain the translated text which is finally taken as output. The tree is in in-order.

5. The Algorithm

- The algorithm developed for the software is as follows-

- Generate tokens from the sentence based on white space

- Count the number of tokens to generate loop

- Send the tokens to language generator
- block, to gather semantic information
- like part of speech, tense, number.

- Analyze the sentence on the basis of
- Subject Verb and Object

- Map the tokens with the target language
- keeping in view the Tense , Aspect and
- Modality of source and the target language

- Generate the output in the target language

6. EXAMPLE

We can take up an example to explain the above mentioned modules step by step by taking a simple sentence in English.
“Ram goes to school with his friend by car.” which would be translated to “*रामः स्व मित्रेण् सह् कारयानेन् विधालयम् ग्च्छति*”.
The first process will take the sentence as input and pass it to the sentence analyzer process. This process will generate the tokens for e.g. ram, goes, to, school, with, his, friend, by, car. The next step would be morphological analysis it will provide grammatical information related to the tokens and will extract the root words from the tokens.
On doing Noun morphology on the ram we will get ram (root word) likewise on doing verb morphology on goes we will get go (root word). While mapping this information will play a vital role as in case of the target language the words carry in built information on grammatical features like number, tense, preposition, conjunction etc.
The parsing process will check whether the sentence is grammatically correct or not. The parse tree generated will be passed on to the mapping process which will map source tokens with the target ones which will be a one to one mapping .This information will be passed to the morphological analyser from where the information on root words and the required information would be added and final output would be generated, as shown in Fig. 9 and 10 [5].

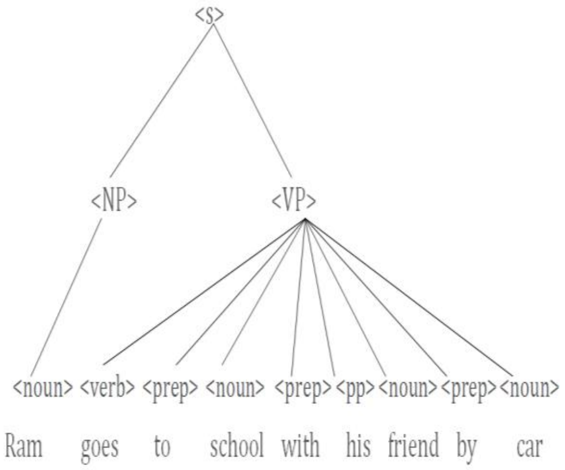


Fig 9: Tree-1 generated in source langauge

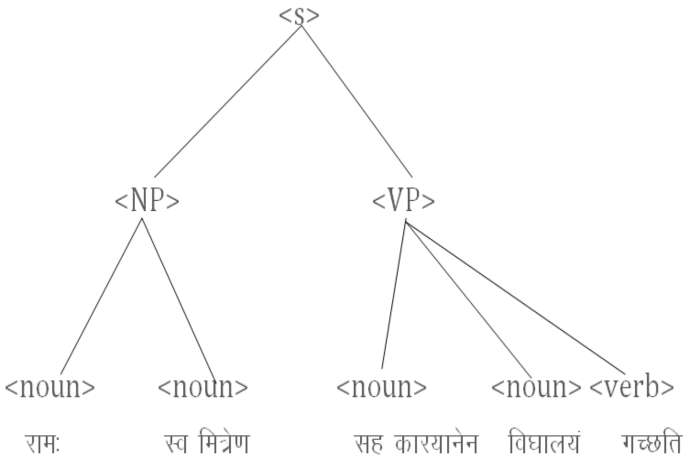
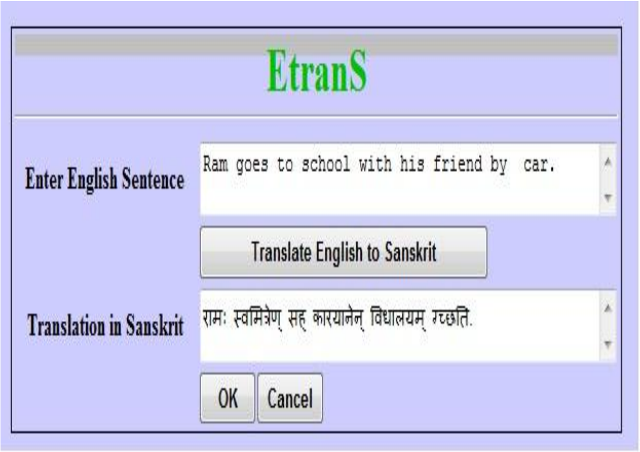


Fig 10:TREE-2[2] generated in target langauge


EtranS

Enter English Sentence: Ram goes to school with his friend by car.

Translation in Sanskrit: रामः स्वमित्रेण् सह् कारयानेन् विधालयम् ग्च्छति.

FIG 11: A SAMPLE SNAPSHOT OF ETRANS

[2] Tree 2shows that “his friend” is replaced by “स्व मित्रेण”, “by car” is replaced by “सह कारयानेन” and “to school” is replaced by “विघालयं”.

7. RESULT

The EtranS system took samples of near about three hundred sentences. The sample snapshot is shown in Fig. 11. The sentences are in active voice, divided into three tenses, simple and compound in nature and having affirmative and imperative types. The sentences are divided into three categories that are small, large and extra large to find out the accuracy of the EtranS system. The performance of translation is graded as three categories A, B and C given as below:

Category	Description	Remarks
A	Sentence is correct in terms of grammar and translation.	-NIL-
B	Sentence is correct in terms of grammar but translation is not correct.	Due to the linguistic representations, few words in English may have multiple roles to play, e.g., the word became is used as a multipurpose word in English, e.g., He became king. She became sad. In Sanskrit, there are different representation for became in the above example. This is a constraint for the software but a linguist can decide where to use which word.
C	Sentence is ambiguous, i.e., it failed at the parsing level.	Few words in English may be used as both noun and verb. This generates ambiguity for the system. For example “Leaves are falling from the tree”. “Train leaves at two p.m”. A further line of work is required in this area to understand these anomalies.

Table 3: Categories used for result analysis

Sentence	A (%)	B (%)	C (%)
Small	99	1	0
Large	95	4	1
Extra Large	90	8	2

Table 4: Analysis of the EtranS system

8. CONCLUSION

In this paper, the complete framework for Rule Based Translation is outlined. The chosen language pair is English and Sanskrit, as a source and target language. The system (EtranS) supports both English and Sanskrit grammar such as noun, verb adjective etc. The system translates simple to compound active voice sentences in English to Sanskrit. It is our belief that this methodology can be adopted for translation of similar languages. The rule base can be extended to translate various types of literature in English to Sanskrit. In the proposed approach we have obtained ninety-nine percent of correctness for the small sentences and ninety percent accuracy for the extra large sentences. The result shows ninety percent of the sentences are correctly translated, however due to linguistic ambiguities two percent of the sentences have reported error.

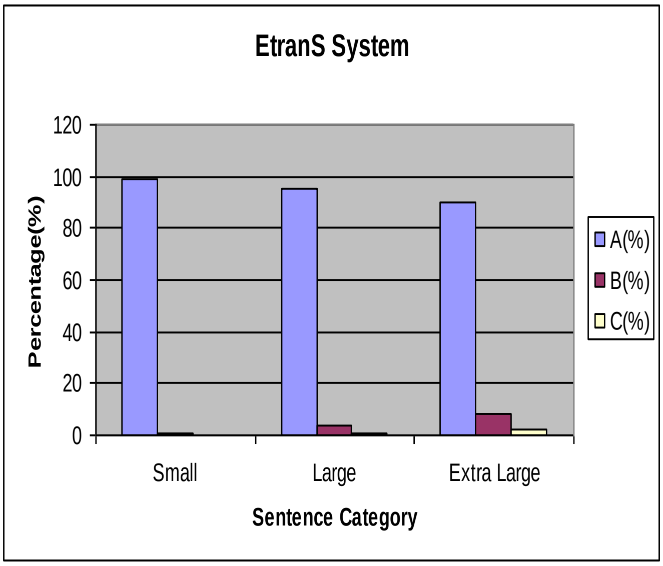


Fig 12: Performance of the EtranS system

8. REFERENCES

[1] English to Sanskrit Machine Translation; Promila Bahadur, A.K Jain,D.S Chauhan, ACM Digital Library, 2011, ISBN: 978-1-4503-0449-8,10.1145/1980022.1980161

[2]Shachi Dave etal, Interlingua-based English-Hindi Machine Translation And Language Divergence, Machine Translation, Volume 16, Issue 4, Pages: 251 - 304, 2001.

[3] James Allen, Natural Language Processing, Pearson Educations

[4] Anurag Seetha etal, Improving performance of English-Hindi CLIR System using Linguistic Tools and Techniques, Pages: 261 HCI 2009.

[5] GB Theory Based English to Hindi Machine Translation System Alka Choudhary, Manjeet Singh, and IEEE