ISSN 2395-1621

Marathi to English Machine Translation for Simple Sentences

[#1]Adesh Gupta, [#2]Aishwarya Desai ,[#3]Nikhil Mehta,[#4]G V Garje,

[#1]adesh1993@gmail.com
[#2]aishwarya.desai93@gmail.com
[#3]nikhilmehta1901@gmail.com
[#4]garjegv@yahoo.com

[#123]PVG's College of Engineering and Technology,Savitribai Phule Pune University Pune, India.

ABSTRACT

Marathi is a regional Indian language and consists of a lot of literature that could be useful if projected in the universal English language. As manual translation is a tedious task, we propose a machine translation system that translates simple Marathi sentences to English using a rule based approach. This approach produces better quality translation than other approaches like statistical which is used by Google for its translation system.

Keywords— Natural Language Processing, Rule-based Machine Translation, Marathi, English, Grammar.

ARTICLE INFO

Article History

Received : 5th June 2015

Received in revised form :

6th June 2015

Accepted : 9th june 2015

Published online :

11th June 2015

I. INTRODUCTION

Translation process is an extremely complex process and challenging, and requires an in-depth knowledge about grammar of both the languages i.e. Source language and Target language to frame the rules for target language generation. Marathi is one of the top 22 Indian languages[2]. Also, about 1% of the world's population speaks Marathi [2]. Translation of Marathi to English will be very useful since English has a global reach. The Marathi to English translation system has numerous applications such as tourism, health care, education, government circulars, medical, insurance etc. Manual language translation is extremely time consuming and costly. The work of Marathi to English translation is in its early stage. The machine translation systems so far developed for any language pair can produce a translation in target language which can give a gist of meaning but may not be able to give exact meaning of a source language sentence or paragraph. Moreover, the translation may change person to person and may have ambiguous words. Some of the Machine Translation systems provide a facility of manual post-processing to select appropriate translation amongst list of translations produced by the system. A Rule-based Machine Translation approach is proposed in this paper to develop a Machine Translation system for Marathi to English to translate simple Marathi sentences.

II. RELATED WORK

Machine translation is a vast topic and many people have been working in this research area for quite some time now. A little work is done for Marathi to English Machine translation. Google has released a beta version of Google Translate [3] to translate Marathi sentences to English using statistical machine translation approach and we have considered it as a reference machine translation system to compare the results. Statistical Machine translation treats problem of translation as a machine learning problem, i.e. the system examines many human produced translations and learns to translate. Statistical Machine Translation uses a mathematical model for achieving translations. The basic requirement is a bilingual corpus (collection of sentences for both source language and the target language). The Statistical Machine Translation system analyses the corpora and by using this probabilistic analysis of bilingual corpora and selects the highest probability translation[9].

III. SYSTEM ARCHITECTURE

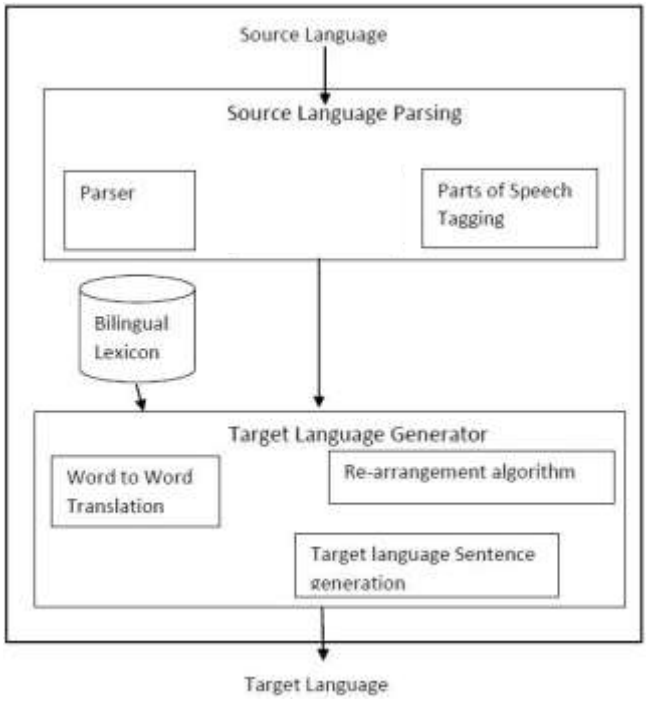


Figure 1: System Architecture

The system architecture is as shown above. It consists of the following components [7].

A. Source Language Parsing
  a) Parsing
  b) POS Tagging
B. Bilingual Lexicon
C. Target Language Generator
  a) Word to Word Translation
  b) Re-arrangement Algorithm
  c) Target Language Sentence Generation

A. Source Language Parsing:

Source language parsing is performed using three components: Parser, Named Entity Recognizer and Parts of Speech Tagger. The parser processes the input sentence and separates each word. Named Entity Recognizer associates each word with its root word. This makes the translation and target language word matching easier. Parts of Speech tagger tags each word with its role in the sentence, e.g. a word maybe a noun, verb, adjective, etc. The output of the source language parsing is passed to the Target Language Generator.

B. Bilingual Lexicon:

A bilingual lexicon is used for matching words from source language to the target language and also for target language sentence generation. It contains association of source language words with the target language words. The source language words are searched in the lexicon based on the root words provided by the parser. Then, the corresponding target language word is retrieved and inflections are added to it, to make its meaning equal in context with the source language sentence. A rule based approach is followed [1].

This lexicon has been manually built for around 4000 words in English. The lexicon is categorized just like a dictionary in the xml format. It consists of dictionary entries as English words and their corresponding Marathi words. The words even have their morphology i.e. morphological as well as semantic properties to define that word.

C. Target Language Generator:

Target language generator is implemented using three components: Word to Word Translator, Re-arrangement Algorithm and Target Language Sentence Generator. The Word to Word Translator translates the source language words into target language words using the Bilingual Lexicon. Re-arrangement Algorithm then re-arranges these target language words into the correct target language sentence structure. The Target Language Generator takes this output and displays the sentence into the target language.

For target language generation, a database is prepared for the grammatical rules of the source language and target language. The database consists of a sequence of *lexical categories* for Marathi language which are mapped to its corresponding English language sequence, which is to be used in the target grammar generator. When a specific set is queried by the Target Language Generator the rules database returns a specific *sequence* to be used for accurate translation after rearrangement of words.

IV. THE PARSER

by IIIT, Hyderabad, India. It provides the system with the morphological analysis of a Marathi sentence.

The Parser provides output in Shakti Standard Format [4] [5]. It provides the root word, tense, gender, multiplicity, direct or oblique case, suffix, vibhakti and other details important to identify the role of the word in the sentence.

The output is represented as a sequence of abbreviated features, with each feature having a fixed position and meaning. These eight cases are mandatory for the morph output:

<fsaf = 'root,lcat,gend,num,pers,case,vibh,suff' >

- **Root**- indicates the root word of the word morphed
- **Lcat**- gives the lexical category of the word. The values it can take are: Noun (n), pronoun (pn), verb (v), adjective (adj), adverb (adv), number (num), etc.
- **Gend**- gives the gender of the word in context. The values it can take are: male (m), female (f), neutral (n).
- **Num**-gives the impression of the word being singular or plural in nature. The values it can take are singular (sg), plural (pl), any.
- **Pers**-gives whether the speech of the word is in the first person (1), second person (2) or the third person (3).
- **Case**-gives whether the noun has a direct or an oblique case depending on the sentence and usage.
- **Vibh**-is the vibhakti of the word.

- **Suff**-identifies the suffix of the word if it contains any.

Example:
हरिहरेश्वरला NNP <fs af='हरिहरेश,unk,,,,,वर#ला,वर_ला' poslcat="NM">

V. DATABASES CREATED (OR USED)

The databases used in the system are as follows:

A. Bilingual Lexicon

This lexicon has been manually built for around 4000 words in English. The lexicon is categorized just like a dictionary in the xml format. It consists of dictionary entries as English words and their corresponding Marathi words. The words even have their morphology i.e. morphological as well as semantic properties to define that word.

B. Database of Tourism Domain

The one of the challenge for Machine Translation is the unavailability of test database or corpus. The bilingual corpus for tourism domain is prepared by TDIL (Technology Development for Indian Languages) is made available for researchers. We have taken a subset of simple sentences from this corpus to test our system [21] [22]. It consists of domain specific meanings which can be considered if ambiguity occurs.

C. Rules Database

This is a database is prepared for the grammatical rules of the source language and target language. The database consists of a sequence of lexical categories for Marathi language which are mapped to its corresponding English language sequence, which is to be used in the target grammar generator. When a specific set is queried by the target language grammar generator the rules database returns a specific sequence to be used for translation after rearrangement of words. XML files have been used to store and maintain the databases due to easy parsing techniques as provided by java.

VI. ALGORITHMS FOR TARGET LANGUAGE GENERATION

D. a Suffix Handling

Suffixes in Marathi get converted into prefixes in English usually barring some exceptions.

For example-
छत्री**खाली** -here root word is छत्री and suffix is खाली.
This suffix becomes prefix during translation
Hence छत्री**खाली**→**under** umbrella
The following table gives an idea of suffix handling. The main suffixes are listed out below

Table A: Suffix handling

Suffix(for source language)	Prefix(for Target language)
पासून	From
साठी	For
तून	From/in
वर	Over/on
मध्ये	In
खाली	Below
चा, चे ,ची, च्या	Other Features
हून	
मागे	Behind
समोर	Front of
पुढे	Front of/ahead of
त	To
ही	Also
स	To/for
ला	To/for

Along with this there are many more suffixes that are handled.

E. Past Tense Handling Algorithm

This algorithm is used for handling the past tense. Whenever the tense is past tense, the word changes a little and is handled by this algorithm.

```
If the tense is tagged as past
        Call handle_past_v() function

handle_past_v()
start
        if past tense then
                find the verb and change its tense to past
                        add  "ed "
                for special cases
                        find the correct past representation of the word
                        assign that to the verb's index in the array of English words
end
```

F. Handling Singular/Plural words

Attaching's' to a word in English is not straight forward. It depends on the last or last but one letter of each word [10]
Example - boy→ boy**s**
Knife→kni**ves**

Rules for attaching 's'

Table B: Attaching s to a word

Last letter of a word	Attach
A,B,C,D,E,F,G,I,J,K,L,M,N,P,Q,R,S,T,U,W,Y	S
H,O,S,X,Z	Es

Exceptions-

i) Consonant + y → consonant + ies
ii) vowel + y→vowel + y + s
iii) alphabets + f→alphabets + ves
Many exceptions are to be handled by coding -

Example- tooth→teeth
Man→men
Nucleus→nuclei
Mouse→mice etc.

VII. WORKING OF THE TRANSLATION SYSTEM

The system implemented using the architecture depicted in figure 1. The features of the system include a rules database along with a lexicon. It also involves disambiguation of prepositions for some rules. All the components implemented in the system are explained with an example in the following section.

Example: अमेरिकेचेचलनडॉलरआहे

A. Parser

Parser provides output in the Shakti standard Format. The output for the given example will be:

<Sentence id="1">
1 ((NP <fs af='अमेरिका,n,f,sg,,o,चे,ं_चे' poslcat="NM" head="अमेरिकेचे">
1.1 अमेरिकेचे NNP <fs af='अमेरिका,n,f,sg,,o,चे,ं_चे' poslcat="NM" name="अमेरिकेचे">
))
2 ((NP <fs af='चलन,n,n,sg,,d,,' poslcat="NM" head="चलन">
2.1 चलन NN <fs af='चलन,n,n,sg,,d,,' poslcat="NM" name="चलन">
))
3 ((NP <fs af='डॉलर,n,,,,,,' poslcat="NM" head="डॉलर">
3.1 डॉलर NN <fs af='डॉलर,n,,,,,,' poslcat="NM" name="डॉलर">
))
4 ((VGF <fs af='आहे,v,,,,,,' head="आहे">
4.1 आहे VM <fs af='आहे,v,,,,,,' name="आहे">
))
</Sentence>

B. Word to Word Translator

This is based on a Lexicon containing around 4000 words including root words. The lexicon consists of its corresponding English words. Its output will be:

America currency dollar is

C. Suffix and Plural Handler

The suffix and plural handler add inflections to the obtained English root words so that their meaning and relativity will be easy to understand in the translated sentence. For the test sentence, after suffix and plural handling, the words will be:

America’s currency dollar is

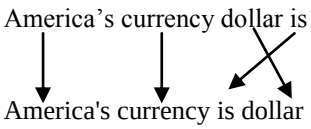
D. Rearrangement Generator

The rearrangement generator provides output in the form of a sequence in which the translated words are to be rearranged according the sentence structureof target language so as to get the output in proper format. The output of rearrangement generator for the test sentence is:

0 1 3 2

E. Target language Grammar Generator

The target language generator will generate the final sentence after rearranging the words in the sequence provided by the rearrangement generator. The output for the test sentence will be:



The words are stored in an array with 0-indexing. Therefore, the Rearrangement Generator’s output will rearrange the words of index: 0 1 2 3 to 0 1 3 2 as shown.

Front end:

VIII. RESULTS

The evaluation tool used to measure the translations quality is BLEU (Bilingual Evaluation Under Study) [6]. This provides a score for a candidate translation compared to a reference translation. The reference translations in our project are translations obtained from linguists who are proficient in English and Marathi. The candidate translation includes translations obtained from our system and an existing system named "Google Translate" developed by Google Inc. The number of sentences on which these results are obtained is around 1020 simple sentences. The following score is obtained on a scale of 0 to 1:

TABLE I
RESULT ANALYSIS

	Our System	Google Translate (by Google Inc.)
Human Translation	0.89	0.654

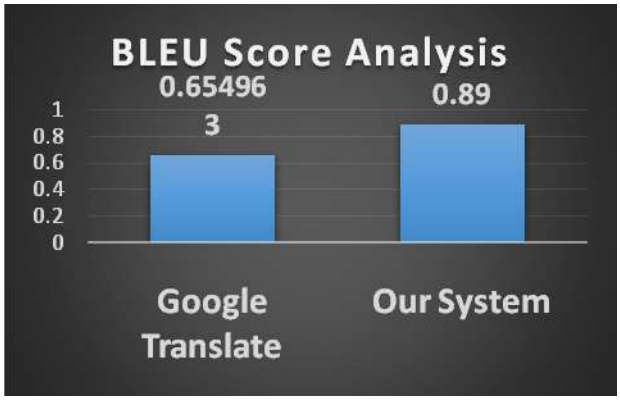


Figure 1: Test Result Bar Chart

A. Test Cases

Some test cases are given below. These test cases have covered most of the tenses and structures. They also cover list processing part. Below are 12sentences, each requiring a different kind of processing.

1) **तीसांगते**
She tells

2) **ऐश्वर्याआंबाखाते**
Aishwarya eats mango

3) **खूपमोठाकिल्लाआहे**
Fort is very big

4) **सातपुरींच्यादर्शनानेमोक्षमिळतो**
Satpuri's darshan gives salvation

5) **जबलपूरसमुद्रसपाटीपासून१३०६फूटउंचीवरआहे**
Jabalpur from sea level is 1306 feet on tall(height)

6) **केरळचीसंस्कृतीहजारोवर्षेजुनीआहे**
Kerala's culture is thousands of years old

7) **गुलमर्गशहरापासुन५२किलोमीटरअंतरावरआहे**
Gulmarg is 52 kilometers on distance from city

8) **परशुरामकुंडएकपवित्रतीर्थस्थळआहे**
Parshuram basin is one holy shrine

9) **गौडचेनमालिकीधबधबाखूपलोकप्रियआहे**
Godchin Maliki waterfall is very popular

10) **पृथ्वीचादुसराप्रमुखखंडआफ्रीकाआहे**
Earth's second major continent is Africa

11) **युरोपमध्येप्रवासाचाखर्चसाधारण२८०००डॉलरआहे**
In Europe travel's expenditure is normally 28000 dollars

12) **येथेतुम्हीनौकाविहाराचाआनंदघेऊशकता**
Here you can take boat riding's pleasure

IX. RESULTS

It has been observed that the rule based machine translation involves generating a lot of rules and handling their exceptions as well. The testing was done on approximately 1020 simple assertive sentences. We are able to achieve far better (24 percent better results as compared to the existing system for our test data) results. As far as the disambiguation part is concerned, the word disambiguation involves a lot of work. Considering all these challenges, we can say that the system is feasible up to a certain extent.

This system can be extended in many ways. The system is now working for simple assertive sentences. So it can be extended for other types of simple sentences such as interrogative, exclamatory etc., as well as complex and compound sentences. The system now works for sentences in tourism domain. Hence it can be implemented for other domains as well because the rules generated are generalized in nature. The system can be also used as a module for a universal system. Apart from these extensions disambiguation of nouns and verbs will be a major improvement to the system.

REFRENCES

[1] Abhay Adapanawar, Anita Garje, Paurnima Thakare, Prajakta Gundawar, Priyanka Kulkarni, "Rule Based English to Marathi Translation of Assertive Sentence " International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013 1754 ISSN 2229-5518
[2]http://www.censusindia.gov.in/(S(22mhid3qsi25vfynyklqv245))/Census_Data_2001/Census_Data_Online/Language/Statement1.aspx Retrieved 28-09-2014.
[3] https://translate.google.co.in/#mr/en/ Retrieved 28-09-2014

[4] http://ltrc.iiit.ac.in/analyzer/marathi/ Retrieved 28-09-2014.

[5] Akshar Bharati, Rajeev Sangal, Dipti M Sharma, "SSF: Shakti Standard Format Guide" (30 September, 2007)

[6] Papineni, K. Roukos, S. Ward, T.; Zhu, W. J. , "BLEU: a method for automatic evaluation of machine translation", *ACL 40th Annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002

[7] Prof. G.V. Garje, Adesh Gupta, Aishwarya Desai, Nikhil Mehta, Apurva Ravetkar, "Marathi to English Machine Translation for Simple Sentences", Volume 3 Issue 11 ,November 2014

[8] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu "Bleu: a Method for Automatic Evaluation of Machine Translation", IBM Research Report, September 17, 2001

[9] Ananthakrishnan Ramanathan, "Statistical Machine Translation", Department of Computer Science and Engineering, Indian Institute of Technology, Bombay

[10]http://www.myenglishpages.com/site_php_files/grammar-lesson-plurals.php

