International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Marathi to English Machine Translation for Simple Sentences

G V Garje[1], Adesh Gupta[2], Aishwarya Desai [3], Nikhil Mehta[4], Apurva Ravetkar[5]

[1]HOD, Department Of Computer Engineering, PVG's College of Engineering and Technology, Savitribai Phule Pune University, Pune, Maharashtra, India

[2,3,4,5] Savitribai Phule Pune University, PVG's College of Engineering and Technology, Pune, Maharashtra, India

Abstract: ***With globalization English has become the official language of the world. With about 71 million Marathi speaking people and varied works in Marathi literature and novels calls for translation. A system is proposed that translates simple Marathi sentences to English using Rule based approach. The system makes use of an online POS (parts-of-speech) tagger maintained by TDIL. Using rule based approach the system is feasible up to certain extent.***

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

1. Introduction

About 71 million of the earth's 7 billion people speak Marathi as their native tongue [3]. Marathi is one of the top 22 official languages of India [6]. Research and other documents in all the fields these days are usually in the English language that are universally recognized and accepted. Existing documents that are presently in the Marathi language need to be translated to English for their widespread use. But, manual translation is costly, time consuming and this give rise to the need of an automated translation system which would do the job in an effective way. Such an automated system developed as a web based or mobile based application makes it suitable for a wide range of use.

2. Challenges

Due to structural difference in source language (Marathi-Subject-Object-Verb) and target language (English–Subject-Verb-Object), there are many challenges in Marathi or Indian languages to English translation. Some of the challenges are listed below [7]:

- Translation accuracy
- Development of generalized translation system
- Unavailability of Lexical Resources
- Difference in methods of encoding information
- Structural Differences
- Lexical Differences
- Case Suffixes
- Verb Related elaborations
- Noun Inflections
- Preposition Disambiguation
- Adjective Inflections

3. Study of Existing Morphological Analysis System

The morphological system that is being used is developed by a consortium of institutions in India which is maintained by IIT Bombay and is funded by TDIL (Technology Development for Indian Languages), Department of Information Technology, Government of India [4]. The system accepts a Marathi sentence/paragraph as input in the UTF-8 or WX format and gives a morphological analysis of the sentence/paragraph in respect to various attributes that help us in identifying the context of the sentence/paragraph. It gives us morphological information such as category, gender, suffix, number, person and root of each word in the sentence. In Marathi, nouns inflect for gender, number and case. To capture their morphological variations, they can be categorized into various paradigms based on their vowel ending, gender, number and case information. The morphemes attached to a verb help identify values for Gender, Number, Person, Tense, Aspect, Modality features for a given verb form. We are using this parser for processing source language [4].

3.1 Attributes

There are various paradigms which are characterized by this system for each word in the given Marathi sentence based on their Part of Speech (POS) usage in that sentence. Verbs inflect for grammatical properties such as gender, number, person, tense, aspect and mood.

- Aspect: Grammatical Aspect of a verb defines the temporal flow in the described event. Different kinds of aspect are Habitual, Perfect, Stative, Completive, Progressive, Durative and Inceptive.
- Mood: Grammatical Mood describes the relationship of a verb with reality and intent. Its various kinds of mood are Subjunctive, Imperative, Abilative, Conditional, Permissive and Optative.
- Tense: Grammatical Tense is a temporal linguistic quality expressing the time at, during, or over which a state or

Licensed Under Creative Commons Attribution CC BY

action denoted by a verb occurs. Tense can be Past, Present or Future.
- Person: Person is the reference to the participant role of a referent, such as the speaker, the addressee, and others. Person can be First, Second or Third.
- Gender: Gender indicates the whether the agreeing noun is masculine, feminine or neutral.
- Number: Number indicates the whether the agreeing noun is singular or plural.

Nouns inflect for gender, number and case. Adjectives and pronouns also inflect for the same.

- Gender: Indicates whether the noun is masculine, feminine or neutral.
- Number: Indicates whether the noun is singular or plural.
- Case: Indicates whether the noun has direct or oblique case depending upon its usage in the sentence.

3.2 Output of Analysis

The analysis of the input Marathi sentence is represented in the Shakti Standard Format (SSF) [5], which makes it easier for computation and also gives us a fixed representation of the analysis so obtained. 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:

<fs af = '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

E.g. For the sentence "मी घरी आहे." We get the parser output as:

```
<Sentence id="1">
1     ((   NP   <fs af='मी,pn,,sg,1,d,,' head="मी">
1.1   मी   PRP  <fs af='मी,pn,,sg,1,d,,' name="मी">
      ))
2     ((   NP   <fs af='घर,n,n,sg,,o,ई,' head="घरी">
2.1   घरी  NN   <fs af='घर,n,n,sg,,o,ई,' name="घरी">
      ))
3     ((   VGF  <fs af='अस,v,,sg,1,,,' aspect=s tense=pre
                mood=in head="आहे">
3.1   आहे VM   <fs af='अस,v,,sg,1,,,' aspect=s tense=pre
                mood=in name="आहे">
3.2   .    SYM  <fs af='.,pun,,,,,,' poslcat="NM">
      ))
</Sentence>
```

The abbreviations can be understood with the help of the following description:

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Figure 1: Tags for Parts of Speech of Parser

4. Proposed System Architecture

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

- Source Language Parsing
- Bilingual Lexicon
- Target Language Generator

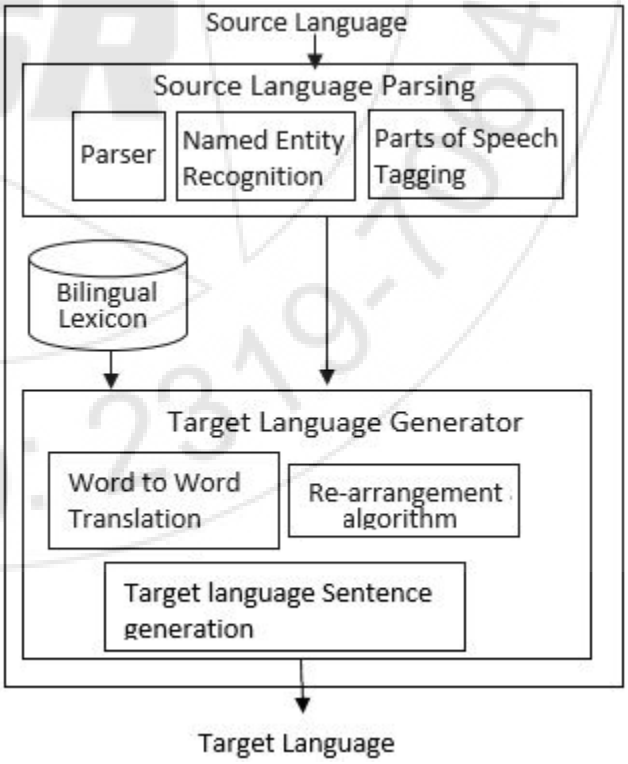

Figure 2: System Architecture

Licensed Under Creative Commons Attribution CC BY

i) Source Language Parsing

Source language parsing is implemented 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 with each word 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.

ii) Bilingual Lexicon

A bilingual lexicon is used for matching words from source language with 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 Named Entity Recognizer and then the variation of the root word in the target language is found by the part of speech the word belongs to. A rule based approach will be followed [1].

iii) 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 converts the Source Language words into Target Language using the Bilingual Lexicon. Re-arrangement Algorithm then rearranges 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.

5. Scope of Use

5.1. Advantages

India is a country with a large population well versed with vernacular languages but not fluent in English. A Marathi to English translation system will be helpful to the Marathi speaking population who need to converse in English. Lot of documents, scripts and scriptures in Marathi also need to be translated to English and this process is manual. Marathi to English translation system will help to automate this process and help reduce manual work related to translation.

5.2. Limitations

Considering the number of rules [2] to be included in the system, it is not possible to achieve perfect translations for each and every sentence. There might be some disambiguation present in some sentence translations. It is also language specific and cannot be used for translation of any other language pair. The testing of the rules will be done for tourism domain because bilingual corpus for this domain is available with TDIL. However rules for translation will be framed in such a way that the general sentences or sentences from other domain will be translated.

5.3. Applications

The system has a wide range of future applications:

- For translation of Marathi manuscripts into English
- Use as an interface for a bigger Translation system
- Extending the systems for other domains

6. Conclusion

In the field of Machine Translation the first generation consisted of dictionary based methods which involved word to word translations. Its shortcomings led to the second generation which involved rule based and transfer based techniques. It has been observed that rule based machine translation involves generating a lot of rules and handling their exceptions as well. The system is feasible up to a certain extent but the translation quality will be better in this method. This paper focuses on rule-based Marathi to English Translation. It can still be said that no such method exists for perfect translations.

References

[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] Rekha Sugandhi, Charugatra Tidke, Shivani Patil, Shital Binayakya ,”Modified Mapping Rules For English To Marathi Translation”, International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 3 Issue 3 (May 2013)

[3] http://www.censusindia.gov.in/(S(22mhid3qsi25vfynyklqv245))/Census_Data_2001/Census_Data_Online/Language/Statement1.aspx 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] G.V. Garje, G.K. Kharate, Minal R. Apsangi, Harshad M. Kulkarni, Manasi S. Sant “Challenges in Rule Based Machine Translation From English To Marathi”, in proceedings of International Conference on Recent Trends in Engineering and Technology (ICRET’14), published in Elsevier digital laboratory.

[7] G.V. Garje, G.K. Kharate, “Survey of Machine Translation Systems in India”, International Journal on Natural Language Computing (IJNLC), October 2013, Vol. 2, No.4, pp. 47-67 Available: http://airccse.org/journal/ijnlc/current2013.html

Licensed Under Creative Commons Attribution CC BY