A NOVEL APPROACH FOR RULE BASED TRANSLATION OF ENGLISH TO MARATHI

Amruta Godase[1] and Sharvari Govilkar[2]

[1]Department of Information Technology (AI & Robotics), PIIT, Mumbai University, India

[2]Department of Computer Engineering, PIIT, Mumbai University, India

ABSTRACT

This paper presents a design for rule-based machine translation system for English to Marathi language pair. The machine translation system will take input script as English sentence and parse with the help of Stanford parser. The Stanford parser will be used for main purposes on the source side processing, in the machine translation system. English to Marathi Bilingual dictionary is going to be created. The system will take the parsed output and separate the source text word by word and searches for their corresponding target words in the bilingual dictionary. The hand coded rules are written for Marathi inflections and also reordering rules are there. After applying the reordering rules, English sentence will be syntactically reordered to suit Marathi language.

KEYWORDS

Syntax analysis, Bilingual, Multilingual, Named Entity Recognition, Word Sense Disambiguation, Morphological Synthesizer, Transliteration

1. INTRODUCTION

This paper presents a novel approach for rule based translator English to Marathi Machine aided translation system. Machine Translation (MT) is the central areas of focus of Natural Language Processing. Machine translation is important for breaking the language barrier among the multilingual country and for facilitating the inter-lingual communication. If we succeed to this, then we can say that exact translation is done by system.

India which is the largest democratic country where more than 30 languages and 2000 dialects used for the communication by the Indians. Because of this different culture and multilingual environment there is a big requirement for translation for the transfer of information and sharing of the ideas, thoughts and facts.

Various MT approaches are exists for developing MT system: 1) Direct based MT 2) Rule based MT 3) Interlingua based MT 4) Statistical based MT 5) Example based MT 6) Knowledge based MT 7) Principle based MT 8) Online Interactive MT 9) Hybrid based MT.

Direct Machine Translation is simplest approach in which a direct word to word translation is done (1). A Rule-Based Machine Translation (RBMT) system includes collection of various rules, a bilingual lexicon or dictionary, and software programs to process the rules (2). Interlingua based approach, this translation consists of two stages, the source Language (SL) which is first converted in to the Interlingua (IL) form a then finally translate into target language. The main advantage of this approach is that the analyzer and parser of SL script is independent of the generator for the Target Language (TL) script and which requires complete resolution of ambiguity in source language text(3). Statistical machine translation (SMT) is a statistical

DOI:10.5121/acii.2015.2401

framework which is based on the knowledge and statistical models which are extracted from bilingual corpora and this is a data oriented structure (4). Basic idea of example based MT is to reuse the examples of already existing translations (5). Knowledge-Based Machine Translation (KBMT) is closely related to Interlingua approach and which requires complete understanding of the source text prior to the translation into the target text. KBMT is implemented on the Interlingua architecture (6). Principle-Based Machine Translation (PBMT) Systems are totally based on the Principles & Parameters Theory of Chomsky‘s Generative Grammar and which formally applies parsing method. In this, the parser generates a tree which shows detailed syntactic structure along with lexical, phrasal, grammatical information (7). In online interactive translation system, the user has full rights to give suggestion for the correct translation which is very advantageous for improving the performance of MT system. This approach is very useful, where the context of a word is not that much clear or unambiguous and where multiple possible meanings for a particular word (8). By combining the advantages of statistical framework and rule-based MT methodologies, a new approach was emerged, which is namely called as “hybrid-based approach”. The hybrid approach used in a number of different ways (9).

This paper is organized into 4 sections. Section 1 discuss an introduction of MT, Section 2 gives brief idea of major MT systems related work in India in tabular format; section 3 introduces the proposed approach to build a MT systems and finally we conclude the paper in the next section.

2. RELATED WORK & LITERATURE SURVEY

In this section we look at some major Machine translation systems of India. Most of the researchers concentrate on Rule based approach because Rule based approach is an easy to build and which is always extensible and maintainable. English to Devnagari Translation is done by M.L.Dhore [1]. The author proposes a hybrid approach and system is specifically developed only for Banking Domain. System translates User Interface labels of commercial web based interactive applications. Devika P, Sayli W. presents a MT system which translates an English sentence to Marathi sentences of equivalent meaning [2]. Abhay A, Anuja G. dealing with rule based translation of assertive sentences [3]. In this system author going through various processes. A novel approach for Interlingual example based translation is developed by K.Balerao, V.Wadne [4]. Transmuter MT is developed for Tourism domain by G. Gajre [5]. ANUVAADAK MT [6] has been hosted online for public access by IIIT Bombay. The System enables translation between different Indian Languages and also provides transliteration support for input of system. SAAKAVA [7] is the websites which carries out translation of an English sentence into Marathi. They are now developing a computer programme with the help of certain dictionary and will try to understand English sentence and then translates the same sentences into Marathi by applying all the rules of Marathi grammar. Google translate is a multilingual service which translate written text from one language to another [8]. It supports 90 languages and many more. The Google translation algorithm is based on statistical analysis and largely depends on a solid corpus.

3. SYSTEM OVERVIEW

Like translation done by human, MT does not simply substituting words in one language for another, but the complex linguistic knowledge; morphology (how words are built from smaller units of meaning), syntax(grammar), semantics(meanings) and understanding of concepts such as ambiguity. The translation process stated as:

1. Decoding the meaning of the source text and
2. Re-encoding this meaning in the target language.

The idea is to translate an input document by going through various phases such as pre-processing, syntax, semantic and lexical phases and finally translating the documents into target language using various mapping rules. The input to the system is a single text document in English Natural Language (NL) and output will be a translated in Marathi NL. The proposed approach consists of 3 phases:

Pre-processing phase, Transfer & generation phase and Post-processing phase. Following Diagram shows the proposed approach.

Algorithm:
Input: Accept a digital document as input in English NL.
Output: Translate document in Marathi NL.

1. Accept a text file as Input.
2. For each sentence in input document do,
3. Apply POS tagging & generate the parse tree for each sentence then,
4. Apply NER rules on each sentence.
5. Perform WSD on lemmas to understand the exact meaning of the lemmas.
6. Use a bilingual dictionary to obtain appropriate translation and transliteration of lemmas.
7. Obtain the proper form of words using Inflections.
8. Represents the sentence based on target language grammar rules.

3.1 Pre-processing Module:

This is the first phase of any machine translation process. This phase is about to make MT process easier and qualitative. The source text may contains figures, diagrams, formulas, flowchart etc. that do not require any translation. So only translation portion should be identified here. It consists of 3 main processes: Syntax analysis, Named Entity Recognition and Word sense Disambiguation.

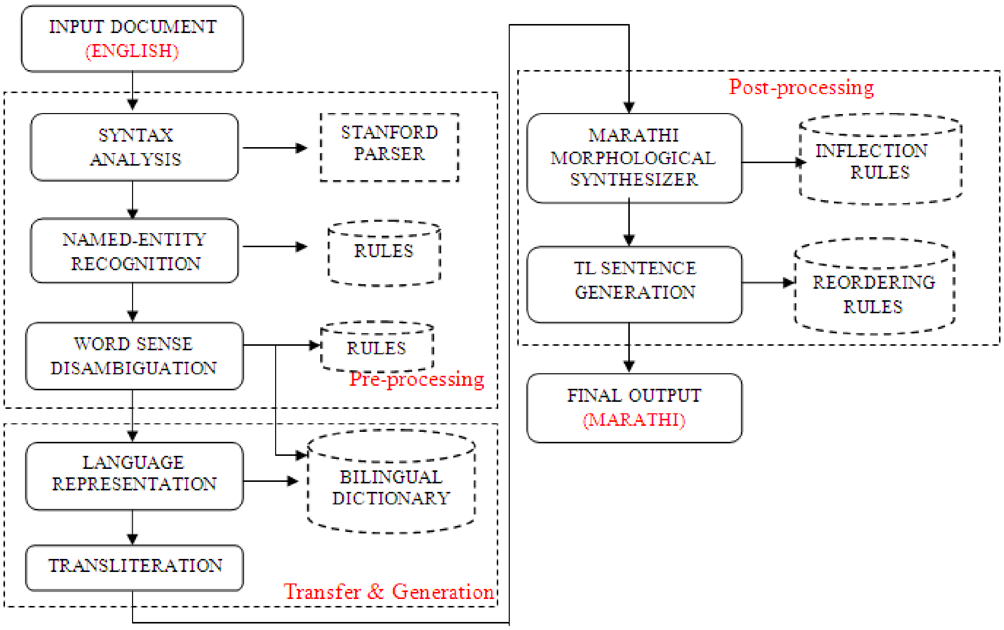

Figure 3.1 Proposed Approach

3.1.1 Syntax Analysis

Syntax analysis exploits the result of morphological analysis to build a structural representation of a sentence. Parser is an algorithm which developed a syntactic structure like tree for a given input. Parser is used for 4 main purposes: To give the parse tree structure of sentences, for Part-of-speech (POS) tagging of English sentences, for stemming the words of English sentences and for chunking of words.

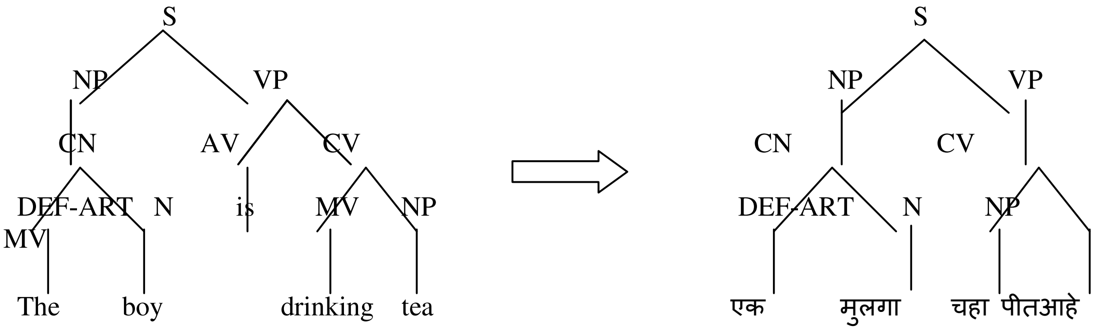

Figure 3.2 English to Marathi Translation of *The boy is drinking tea.*

3.1.2 Named Entity Recognition

Named Entity Recognition (NER) gives sequences of words in a text which are the names of things. It comes with well-engineered feature extractors for Named Entity and for defining feature extractors. Stanford NER tool and Open NLP tool are available for doing the tasks. Various rules are exist for Named entity Recognition:

I) Rules for creating Person's Name

a) Look for Proper Nouns.
b) Contextual words like {men, books, author of, co-author, read, worked, state, city, country, university, college, school, island of, hero, hospital, born, establish, started, saints, founded, chairman of , director} if came then it will consider as proper noun.
c) If set of capitalized word include a set of letters followed by (.), followed by mostly one (rarely two) capitalized words, then the whole set is considered as name.
d) If one of the capitalized words appears subsequently, the probability for it belongs to name.
e) If the set of words or one of capitalized words appear at the beginning of a sentence, it will considered as name.
f) If preposition belongs to {by, of, friend, colleagues, to, co-author, with, men, persons, emperor, men like, sage, as}, the probability for it to be name increases.
g) If the word immediately after the capitalized word(s) (i.e. the post-position) is belongs to set {said, told} the probability for it to be name increases.
h) An apostrophe's ('s) to a capitalized word, then the probability it consider as name.

II) Rules for creating Place /Institute /Organization Name Index

a) Look for Proper Nouns.
b) If a Preposition comes immediately after a Name, it is likely to be a Place or Organization or Institute.
c) Possible set of preposition for potential Place or Organization {from, in, at, to, for, of}

III) Rules for creating Date Index

a) Look for Proper Nouns.
b) Words like Century, decade, AD., BC., during, before, after, until if came in sentences then probably it's a year.

3.1.3 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is defined as assigning the correct sense to a word according to its context. The disambiguation is done by applying construct which is called as restriction. For example,

back bencher => student who is not serious in studies and always sitting at back of the class.

This additional sense will be included in the dictionary with the help of some rule patterns.

The Verb **get off**

Main Word	**Multiple Meaning**	**Meaning**	**Examples**
get off	leave	प्रस्थान करणे	We got off after breakfast
get off	be saved	वाचणे	lucky to get off with a scar only
get off	send	पाठवणे	Get these parcels off by the first post
get off	stop	बंद करणे	get off the subject of alcoholism
get off	stop with obj>work	काम थांबवणे	get off (work) early tomorrow

The Noun **Shadow**

Main Word	**Multiple Meaning**	**Meaning**	**Examples**
shadow	darkness	अंधार	the place was now in shadow
shadow	patch	काळे डाग	shadows under the eyes
shadow	atmosphere	छायेत	country in the shadow of war
shadow	hint	संकेत	the shadow of things to come
shadow	close company	सावली	the child was a shadow of her mother
shadow	deterrent	छाया	a shadow over his happiness
shadow	ghost	भूत	seeing shadows at night

3.2 Transfer & Generation Module:

This is the second phase of machine translation system. This module composes the meaning representation and assigns them the linguistic inputs. Here exact word mapping and transliteration is done with the help of dictionary.

3.2.1 Language Representation

It is necessary to have dictionary. The bilingual dictionary is use for the purpose of a lexicon by storing root words together along with their morphological properties and meaning of English words in Marathi, equivalent synonyms. Dictionary database is endless. Therefore we can extend it according to our need.

The pre-processing of dictionary undergoes various stages depending on the data: 1) Font-converting 2) Aligning 3) POS tagging 4) Adding Synonyms.

Each database has five fields: source, target, category and Synonym.

i. The field ‘source’ stores the English words.
ii. The field ‘target’ stores the Marathi words.
iii. The field ‘POS category’ stores the Part-of-Speech category of the source and target words.
iv. The field ‘synonym’ stores the synonym of Marathi words for that particular English word.

Table 3.1: Typical Entries for English words in a Sample Bi-lingual Dictionary

Source	POS category	Target	Equivalent Synonyms
after	adjective	नंतर	काही वेळाने
abed	adverb	बिछान्यावर	अंथरुणावर
book	noun	पुस्तक	व्यवस्था करणे
across	preposition	आरपार	पलीकडे

3.2.2 Transliteration

Transliteration involves separating tokens form the input query and mapping the characters of each token to its equivalent Marathi characters.

Algorithm:

1) Once the English sentence is entered , the tokens/words from the query are separated as below
    a) From the start of string, identify all the vowels positions one by one till the end of string.
    b) If vowel position is found then the word before the vowel is identified as a token
2) For each token identified in the above step:
    a) For each character in the token, its equivalent English character(s) are mapped.
    b) We have used below mapping between English alphabets with its equivalent Marathi alphabets.
    c) Once for a complete English string, the tokens are identified and transliterated then the transliterated text will be provided as input to next module.

For eg. Shivaji (Sh +i) + (v+a)+(j+i) → शि + वा + जी → शिवाजी

3.3 Post-processing Module:

This is the last and main phase of any machine translation process. Once the text is represented in target language, the target text is to be reformatted after post-editing. Post-editing is done to make sure that the quality of translation is upto the mark. Post-editing should continue till the MT system reach human like. It consists of two main sub-modules.

3.3.1 Marathi Morphological Synthesizer

The language which follows SOV structure use postpositions. Hence while translating an English sentence (SVO structure) to a Marathi sentence (SOV structure), we need to change the position that is we have to convert preposition to postposition.

Words can be classified in two types depending on inflections [6]:

Inflectional Words: Noun, Pronoun, Adjective, Verb

Non-Inflectional Words: Adverb, Interjection & Conjunction

The words are inflected base on the changes in Gender i.e. Masculine, Feminine, Neuter or Multiplicity i.e. Singular, Plural or Tense i.e. Present, past, Future and case.

For Example: Shivaji came to house to meet Jijabai but she was not there.

English Word	Marathi Lemma	Inflected Words	Rules to get inflected
Shivaji	शिवाजी	शिवाजी	------
Came	येणे	आला	Verb, 3rd person, Masculine, Singular, Past tense
to	ला		
house	घर	घरी	Case table, Locative, Singular
to	ला	भेटायला	Case table, Dative Singular
meet	भेटणे		
Jijabai	जिजाबाई	जिजाबाईला	Case table, Accusative Singular
but	पण	पण	-----
she	ती	ती	-----
was	होती	नव्हती	-----
not	नाही		
there	तिथे	तिथे	-----

So for implementation of the inflection we need to store the following information to the database.

English words, POS Tag, Gender, Tense, Multiplicity, Degree and Case.

3.3.2 Target Language Sentence Generation

To obtain a Marathi output from input English sentences, the differences in the syntactic structure must be determined using rules i.e. mapping rules. A set of mapping rules in terms of syntactic structure can be used for English-Marathi MT. Following table shows the reordering rules for English-Marathi translation where, there are very few rules presented to give the idea of how the production rules work.

Table 3.2 Reordering Rules

English Patterns		**Marathi Patterns**	
R1	S → NNP+VBD+DT+NN+IN+ NNP’+NNP”+IN’+NNP”’ Shivaji was the founder of Maratha empire in India.	R1’	S → NNP+(IN’+NNP’”)+NNP’+ (NNP”+IN)+(DT+NN)+VBD शिवाजी भारतात मराठा साम्राज्याचे संस्थापक होते.
R2	S → NNP+VBD+VBN+IN+ CD+IN’+NNP’ Shivaji was born in 1627 at Poona.	R2’	S → NNP+ VBN+(NNP’+IN’)+ (CD+IN)+VBD शिवाजींचा जन्म पुणे येथे 1627 मध्ये झाला.
R3	S → PRP+NN+NNP+NNP’+ VBD+DT+NN’ His father Shahaji Bhonsle was a Jagirdar	R3’	S → PRP+NN+NNP+NNP’+DT+ NN’+VBD त्यांचे वडील शहाजी भोसले एक जहागीरदार होते
R4	S → NNP+VBD+IN+NNP’ Shivaji lived at Bijapur	R4’	S → NNP+NNP’+IN+VBD शिवाजी विजापूर येथे राहत होते
R5	S → PRP+NN+NNP+NNP’+VBD+ DT+RB+JJ+NN’ His mother Jija Bai was a very pious lady	R5’	S → PRP+NN+NNP+NNPJJ+NN’+ VBD त्यांची आई जिजा बाई एक अतिशय धार्मिक वृत्तीची महिला होती.
R6	S → NN+VBD+VBG+NNS Seema was peeling potatoes	R6’	S → NN+NNS+VBG+VBD सीमा बटाटे सोलत होती
R7	S → NNP+VBZ+DT+NN+TO+ VB+TO’+NN’ Knowledge is the way to go to heaven	R7’	S → NNP+(TO’+NN’)+(TO+VB)+ (DT+NN)+VBZ ज्ञान स्वर्गाकडे जाण्याचा रास्ता आहे
R8	S → PRP+VBZ+DT+JJ+NNP It is a costly pen	R8’	S →PRP+DT+JJ+NNP+VBZ तो एक महाग पेन आहे

4. CONCLUSION

This paper presents a novel approach for English to Marathi language translation. This proposed system follows transfer approach with rule based translation. The main focus revolves around morphological synthesizer. We have presented a complete architecture for this MT system and several algorithms for this system. Proposed system uses sentence construction rules both for English and Marathi language. It uses a Stanford parser for tagging and chunking which eliminates the need of unnecessary computations. English to Marathi dictionary will be used to support efficient translation. The goal of developing such translation system is to make the resources available to everyone.

ACKNOWLEDGEMENT

I wish to express my deep sense of gratitude to our guide Prof. Sharvari Govilkar, for providing me with the idea of this topic and for her invaluable support, guidance and supervision. I would like to thank my Principle and Head of Department for their invaluable support.

REFERENCES

[1] Devika Pishartoy, Priya, Sayli Wandkar (2012) “Exteneding capabilities of English to Marathi machine Translator”

[2] Abhay A, Anita G, Paurnima T, Prajakta G (2013), “Rule based English to Marathi translation of Assertive sentence”

[3] Krushnadeo B, Vinod W, S.V.Phulari, B.S.Kankate (2014), “A novel approach for Interlingual example-based translation of English to Marathi”

[4] G.V.Gajre, G.Kharate, H. Kulkarni (2014), “Transmuter: An approach to Rule-based English to Marathi Machine Translation” (2014)

[5] Pushpak B. Jignashu P. “Interlingua Based English-Hindi Machine Translation and Langauge Divergence”

[6] Charugatra Tidke, Shital B, Shivani P (2013) “Inflection Rules for English to Marathi Machine Translation”

[7] www.cflit.iitb.ac.in/indic-translator/

[8] www.saakava.com/index.aspx

[9] https://translate.google.co.in

Authors

Ms. Amruta Godase is Pursuing M.E. in Information Technology with Specialization in Artificial Intelligence & Robotics from Mumbai University. She received her polytechnic from MSBTE & B.E from Mumbai University.

Sharvari Govilkar is Associate professor in Computer Engineering Department,at PIIT, New Panvel, University of Mumbai, India. She has received her M.E in Computer Engineering from University of Mumbai. Currently She is pursuing her PhD in Information Technology from University of Mumbai. She is having Sixteen years of experience in teaching. Her areas of interest are Text Mining, Natural language processing, System programming & Compiler Design etc.