CHALLENGES IN RULE BASED MACHINE TRANSLATION FROM MARATHI TO ENGLISH

Namrata G Kharate[1], Dr. Varsha H. Patil[2]

[1]Department of Computer Engineering, VIIT, Pune, Maharashtra, India
[2]Head of Department, Department of Computer Engineering, MCOERC, Nashik, Maharashtra, India

ABSTRACT

Machine translation is being carried out by the researchers from quite a long time. However, it is still a dream to materialize flawless Machine Translator and the small numbers of researchers has focussed at translating Marathi Text to English. Perfect Machine Translation Systems have not yet been fully built because the fact that languages differ syntactically as well as morphologically. Majority of the researchers have opted for Statistical Machine translation whereas in this paper we have addressed the challenges of Rule based Machine Translation. The paper describes the major divergences observed in language Marathi and English and many challenges encountered while attempting to build machine translation system form Marathi to English using rule based approach. As there are exceptions to the rules and limit to the feasibility of maintaining knowledgebase, the practical machine translation from Marathi to English is a complex task.

KEYWORDS

NLP; Machine Translation; English; Marathi; grammar.

1. INTRODUCTION

Language is one of the most popular medium of communication and there are many languages used in the world for verbal and written communication. Different languages use different ways to encode information.

There is a need of Translation when the information has to be communicated among the people speaking different languages. Translation is a process of encoding the information from one language and decoding it another language using the rules of target language. This process has been attempted for automation between a few pair of languages since a long time. Though accuracy is not fully achieved in the pair of languages and very less attempt has been observed in regional languages such as Marathi – English as a pair, Machine Translation often produces coarse yet understandable translations.

In this paper, Machine Translation from Marathi to English has been considered for simple assertive sentences along with the challenges in the translation. Marathi is an Indo-Aryan language that has more than 42 identified dialects. English, on the other hand is a West Germanic language. Its origin is in the Anglo-Frisian dialects of North West Germany and the Netherlands[2]English is now considered as a global language, whereas Marathi is a language spoken mainly in the central and Western regions of India. English is spoken as a first language by around 375 million people whereas the number of Marathi speakers is 90 million speakers worldwide. Marathi is the 15th most spoken language in the world and 4th most spoken language in India [6].

Natarajan Meghanathan et al. (Eds) : ACSTY, NATP, ADCOM, ITCSS - 2019
pp. 45-54, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.91005

Many official documents and lot of information these days are available in the Marathi language, especially in a state of Maharashtra. Existing documents that are currently in the Marathi language need to be translated to English for their widespread use. Manual translation is very costly and time consuming and hence there is a need to have an automated translation system which would do the language translation in an effective way. There are major challenges in the process due to the structural difference between Marathi language and English language. English follows Subject-Verb-Object grammar structure, while Marathi language follows Subject-Object-Verb grammar structure, relatively of free word order and has large number of inflections. Hence its translation to English is a challenging task. [5]

Further, Marathi is highly dominated by inflections and case-suffixes. Thus, a rule based machine translation system from Marathi to English would have to take into consideration these differences in the languages. Such a Machine Translation system will not only promote the language on a global scale, but it will also open the gates to the people who are facing problems while translating Marathi to English.

Google translator is only tool available for Marathi to English translation .It uses Statistical Machine Translation that is machine translation in which translation is done using statistical translation models, parameters of which are derived from the analysis of bilingual text corpora. If corresponding word is not found in the text corpora, accurate translation is not obtained. Moreover the Google translate does not check the syntax of the given sentence. [7]

2. RELATED WORK

In the existing literature, the issue of translation divergence for Marathi and English MT has not been exhaustively examined. S. B. Kulkarni [2] discuss syntactic and structural divergence issues in English-Marathi machine translation and the same translation pair is then examined for reverse translation so as to examine the nature of the divergence in each case. R.K.sinha[3] discuss different types of translation divergences in Hindi and English MT. G.V.Garje [4] describes the differences between the languages English and Marathi from a Machine Translation point of view and also encountered challenges while attempting to build a Machine Translation system from English to Marathi using Rule based Machine Translation approach. In this paper we describes the major divergences observed in language Marathi and English and many challenges encountered while attempting to build machine translation system form Marathi to English using rule based approach.

3. SYSTEM ARCHITECTURE

Figure 1.0 depicts the overall architecture of the proposed system. The various components of architecture are Parser, POS tagger and name entity recognition. Initially the input Marathi sentence parsed using Marathi parser that using shallow parser. The words in the tree structure are tagged according to their parts of speech. Name entity recognition tagging is used for Transliteration concept in Machine Translation. It also helps in capitalization of English Sentence. A bilingual lexicon includes Marathi word and its respective English word. The sentence from the pre-translation process is split into words. For each word its parallel English meanings are obtained from the lexicon. The lexicon also contains all possible synonyms of English word. The attributes include base forms for noun inflections and base forms of verbs for conjugations.

In some words of the sentence, obtained from the word by word Translator possesses ambiguities due to presence of multiple meanings. These ambiguities must be resolved to obtain better

translations. Few algorithms have been proposed for Word Sense Disambiguation like Lesk Algorithm [11], Walker Algorithm [12], and HyperLexAlgorithm [11].

In Target Language Generator, word to word translation includes translation and transliteration depending on word's POS tag. While for word to word translation one has to consider all rules like noun inflection, verb inflection and adjective inflection. There are various challenges discussed in this paper which needs to be handled during translation. Each challenge requires different rule to design. The rearrangement generator provides output in the form of a sequence in which the translated words are to be rearranged according the sentence structure of target language so as to get the output in proper format. The target language generator will generate the final sentence after rearranging the words in the sequence provided by the rearrangement generator.

4. CHALLENGES IN TRANSLATION

Rule Based Machine Translation uses grammar to formulate transfer-rules from source language to target language. At times, these grammatical rules may not be formally defined. The transfer rules include rules for word-reordering, disambiguation and grammatical additions in the target language. Formation of transfer-rules in a language pair belonging to distant families is a daunting task.

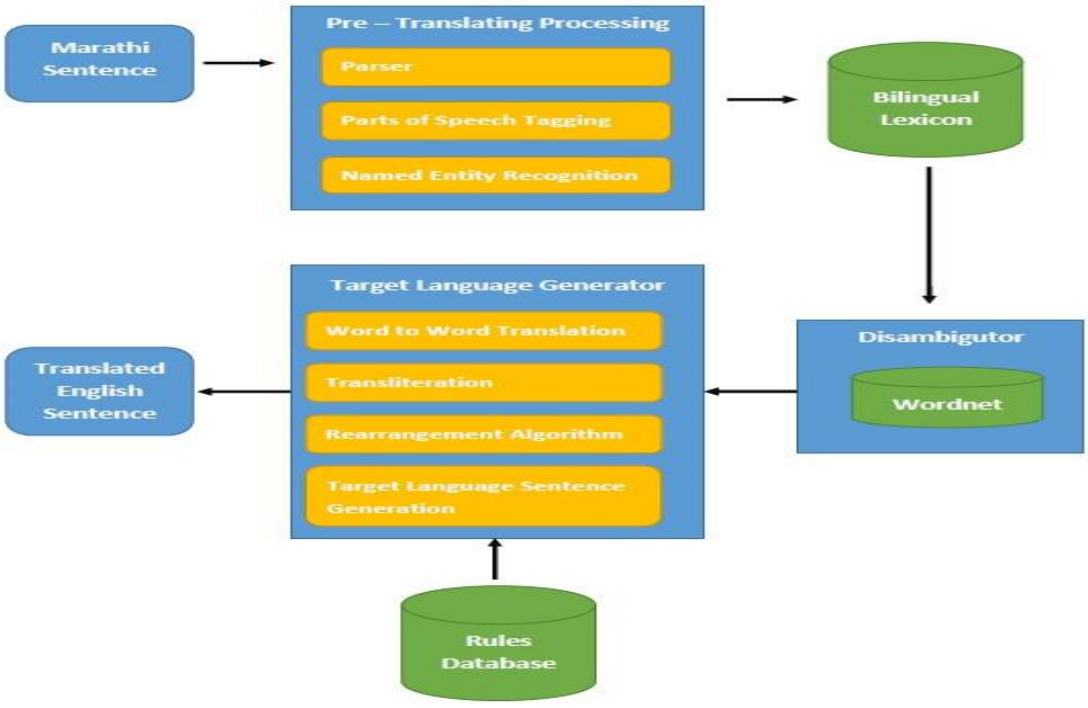

Figure 1. System Architecture

Following are the few major challenges authors have come across.[1][3][4]

1. Unavailability of Lexical Resources
2. Constituent-order Divergence
3. Adjunction Divergence
4. Pleonastic Divergence

5. Case suffix
6. Divergence in Determiner System
7. Replicative Words
8. Expressive Elements
9. Indirect Speech
10. Mapping of Time
11. Difference in methods of encoding information
12. Lexical Gap
13. Adposition
14. Difference in inaminate objects
15. Capitalization
16. Noun Inflection
17. Verb Inflection

1.Unavailability of Lexical Resources

Marathi is a very low resource language [9]. The lexemes in Marathi have their own morphology. It is needed to acquaint with the properties of the language for translation. In high resource languages these may be acquired using language analysis tools like parsers, POS taggers and Named Entity Taggers. The semantic information tools such as language pair dictionaries and wordnets provide with senses of the lexemes. These are later helpful in word sense disambiguation. However these tools are not yet fully developed for Marathi.[4]. So it becomes very difficult to translate as well as disambiguate the words in the sentence. The parallel corpora for Marathi and English are not sufficient to pursue statistical machine translation. Machine translation methods such as Statistical MT and Corpus based MT require a large amount of corpora which are not yet available on a large scale. This poses a restriction to the number of methods which can be used for such a translation system.

2. Constituent-order Divergence

Constituent-order divergence relates to the word-order distinctions between English and Marathi. Essentially, the constituent order describes where the specifier and the complements of a phrase are positioned. For example, in English the complement of a verb is placed after the verb and the specifier of the verb is placed before. Thus English is aSubject-Verb-Object (SVO) language. Marathi, on the other hand, is an Subject-Object- Verb (SOV) language. Example 1 shows the constituent-order divergence between English and Marathi.

Ex.1. तो आंबा खातो आहे.
S O V
He is eating mango.
S V O

3. Adjunction Divergence

Syntactic divergences associated with different types of adjunct structures are classified as Adjunction divergence. Marathi and English differ in the possible positioning of the adjective phrase. In Marathi a Prepositional Phrase(PP)and/or adjective phrase(AP)can be placed between a verb and its object or before the object, while in English it can generally at the terminal of the sentence; consider the following example,

Ex.2. मी उद्या माझ्या बाइकवर आणीन.
S O AP V

I will bring it tomorrow on my bike.
S V O PP

4. Pleonastic Divergence

Another related point of divergence between Marathi and English is regarding the mapping of the words like ‘there’ and ‘it’ in the sentences in English. In English constructions, ‘there’ and ‘it’ are used to denote existential sentences, called as introductory subject. Marathi does not have a pleonastic subject construction and the contrast between existential and non-existential sentences is realized by several other ways such as the movement of the noun phrase from its canonical position and the use of demonstrative elements [1]. Let us consider following sentence.

Ex.3.**खोली मध्ये साप आहे**.
There is a snake in the room.

साप खोली मध्ये आहे.
The snake is in the room.

It is observed that the bare noun phrase साप and ‘snake’ are mapped by indefinite and definite noun phrases in English. However, the only difference between these two Marathi sentences is the respective positions of the subject Noun Phrase(NP) and the खोलीमध्ये adverbial phrase. This type of divergence is related to more than one aspect of grammar such as the word order, lexical and structural gaps in languages. Hence there is a need to examine it in detail to categorize the type of divergence it represents.

5. Case Suffixes

In modern languages there is less number of cases. E.g. in Sanskrit and in Marathi there are 7 cases; in German there are 4 cases while in English there are mainly 2 cases.

Each having its own functional meaning and suffixes. It is difficult to identify these cases from the Marathi sentence and also it is difficult to map cases form Marathi to English. As each case suffix represents different meaning it is utmost important to determine the exact case of the noun. And cases were replaced by prepositions in the evolution of languages. In the absence of case marker the case is called as “Nominative”.

Example.
Second vibhakti is actually preposition "To"
Third vibhakti is preposition "by" as used etc.

6. Determiner System

English has articles that mark the definiteness of the noun phrase overtly. Marathi lacks an overt article system and different devices are used to realize the definiteness of a noun phrase in Marathi. For instance, mapping of a bare NP in Marathi onto an NP with an article “a-an/the” in English is dependent on a detailed syntactic and semantic analysis of the noun phrases in both the languages, as in the following example,

Ex.4**.** मी कुत्रा पाहिला. कुत्रा खूप गोंडस होता
I saw **a** dog. **The** dog was very cute

7. Replicative Words

Marathi has replicative words for which it is difficult to find an exact counterpart in European languages such as English. Almost all kinds of words can be replicated to denote a number of different functions in Marathi. A verb-verb replication such as पाहतापाहता in Marathi cannot be translated as 'see see' in English, even though पाहता and see are the correct translations or adjective replication उंचचउंच can be used to denote different types of functions that are mapped onto English in various ways depending on a number of factors. For example,

Ex.5.**पाहता पाहता सकाळ झाली**.
Ex.6.**येथे उंचचउंच देवदार आणिचीडवृ क्षांशिवाय काहीही नव्हते**.

A closer translation will be 'In the meanwhile, it became morning.' And 'There was nothing here except for dense trees of cedar and pine'.

8.Expressive Elements

Expressive words exist in all natural languages and pose difficulty in processing, particularly in mapping onto another language. The reason is that these words do not have exact parallel translation in another language. Thus the word धडकन/धाडकन is only distantly mapped by "bump" in English, as given below,

Ex.7.ती धाडकन पडली.
She fell with a 'bump'.

The expressive words usually originate from the sound associated with the semantics of the action verb and can be adverbial or verbalized action-verbs. One may argue this to be just a lexical gap but indeed it is not so. However, some of these words can be handled in the lexicon but as in many cases the mapping also involves structural changes, the issue involves a wider scope of interpretation.

9.Indirect Speech

The indirect speech sentences in Marathi and English differ in both the form use of pronominal elements.

Ex.8. राममम्हणाला की मी नाही जाणार.
Ram said that I/he would not go.

The use of the pronoun मीin the example is ambiguous and can be translated either by 'I' or 'he' in English. The example shows that the tense in the English indirect speech sentences is past but must be mapped by present tense in the Marathi sentence. Although some aspects of this type of translation divergence have been partially discussed in Dave et al (2001)[14] for Hindi-English MT, but it needs further examination to comprehend.

10.Mappings of Time

Usually, people's perception of different objects in the world is dependent upon several socio-cultural beliefs. For example, time is conceptualized in the Indian culture differently than that is used in the Western culture. These concepts are expressed through our respective languages and difference in concepts manifests itself in the language that is the source of translation divergence. The representation of small span of the time changes as per the language. The example

belowshows that the time at the 5 o‟clock in the morning is denoted by a.m. in English but the exact translation of ‘sakali’ in Marathi does not produce an appropriate English expression. However, in second example, it is noticed that the time at 3 o‟clock in the afternoon which is expressed in English by p.m. but the exact translation of ‘Duphari’ in Marathi does not produce an appropriate English expression.

Ex.9.तो सकाळी ५ वाजता आला.
He arrived at 5 o‟clock in the morning
He arrived at 5 a.m.

Ex.10.दुपारी ३ वाजता आला.
He arrived at 3o‟clock in the afternoon
He arrived at 3 p.m.

Ex.11.तो संध्याकाळी ७ वाजता आला.
He arrived at 7o‟clock in the evening
He arrived at 7 p.m.

11. Difference in methods of encoding information

English uses word ordering to encode information while Marathi uses morphemes. Here in the Marathi sentence the additions of vibhaktipratyay (case suffixes) ने (ne) and ला (la)conveys that ‘Shyam’ is the doer of the action and ‘Ram’ is the receiver of the action. Even though positions of doer and receiver are different in both sentences, their English translations are same.

For example:
Ex.12. रामला श्यामने आंबा दिला.
Shyam gave a mango to Ram.

Or
Ex.13. श्यामने रामला आंबा दिला.
Shyam gave a mango to Ram.

12. Lexical Gap

English and Marathi are spoken by societies belonging to diverse cultural backgrounds. Hence, there are certain concepts those exist in only one of the languages. Such concepts may cause problems while translating.

For example: ‘कर्म ’ in Marathi.
A solution to this problem is that, some of these concepts are directly borrowed in English.
For example: “karma”

13.Adpositions

Adpositions are words those can occur before or after a phrase, word, or a clause that is necessary tocomplete the meaning of a given sentence. The languages which follow SOV Structure use postpositions. Hence, while translating a Marathi sentence (SOV structure) to an English sentence (SVO structure), we need to change the postpositions (of Marathi) to prepositions (of English). This is a major issue which needs to be resolved for inflecting the nouns, verbs andCases

(Vibhakti). In example below, word ‘var’ appears after noun but in English translated sentence word ‘on’ is before noun.

Ex.14.तो खुर्चीवर बसला
He sat on the chair

14. Difference in inanimate objects

In Marathi gender is an important morphological property of the nouns. In English the inanimate objects are referred using theneuter gender. On the other hand, the inanimate objects in Marathi have their own genders.

For example:
खुर्ची (F) ->Chair (N)
चमचा (M) -> Spoon (N)

15. Capitalization

In the English writing systems that distinguish between the upper and lower case have two parallel sets of letters. While writing sentences in English there are some rules of capitalization for example,the first word of a sentence,The name of people,Months, days, and holidays,The titles of books, articles and moviesetc. Name entity tagger can tag person name, place, city etc., so by using it we can capitalize. In the following example Marathi sentence is written in simple but in English translated sentence first word of sentence and first word of name of state that is Korean is capitalized.

Ex.15. हे विद्यार्थी कोरियन आहेत.
These students are Korean

16. Noun Inflection

Noun inflection is performed on the basis of change in Multiplicity or Case (Vibhakti) in English grammar. The inflection of a word can be determined from the word endings. Noun Inflection in Multiplicity, there are different rules to get plural form of noun .The rules changes according to end alphabets of noun.

Examples.
Box=boxes and Cat=cats
There are some exceptions also for plural forms
Ex. Woman=women

Noun Inflection in case that is possessive case. In possessive case “‘s” is attached to noun, there are different rules to attach “‘s” depending on multiplicity of noun.
Ex. Boys’ school, Men’s club, Boy’s pen

17. Verb Inflection

In Marathi, however, the word according to which the verb gets conjugated is either the doer of the action (karta) or the receiver of the action (karma). In English the main verb in the sentence is conjugated according to the person, tense, number of the subject of the sentence Further, the attributes of the word that determine the conjugation are also different in English.
English verbs consider person, number, tense [9].

For example:
मी क्रिकेट खेळतो ->I play cricket. (First person)
तो क्रिकेट खेळतो->He plays cricket. (Third person)

According to suffix attached with Marathi verb and verb followedआहे/होते/असेल, thereis concept of auxiliary verb in English language. According to tense we have to change to be form and verb form.

For Example
मी खोका उघडला होता -> I had opened box.
तीने दरवाजा उघडला असेल ->She will have opened door.

5. CONCLUSIONS

The translation divergence is a challenging problem in the area of machine translation. A detailed study of divergence issues in machine translation is required for their proper interpretation and detection. Rule Based Machine Translation is a tedious approach. It needs lots of human work to design number of rules. As number of rules increases the quality of translation improves. However, it usually generates nearly accurate translations. As seen in this paper a number of rules need to be formed to achieve these translations. Moreover, resolution of these problems is a pre-requisite for designing a robust machine translation system between the languages considered for the present study. In this paper we have explained the various types of divergence Patterns with respect to Marathi and English language pair. Also the identification and classification of these divergence patterns along with. In spite of all these efforts there exist many exceptions in the language which do not conform to these rules. The accuracy of this approach is dependent on the size of the grammatical knowledge base. Handling of exceptional cases leads to an increase in the size and the complexity of the knowledge base. As the size of the knowledge base increases the accuracy increases up to a certain threshold. If the size increases above certain threshold then the accuracy may reduce due to conflicting rules. Thus a system based on this approach has to balance accuracy and the number of exceptions it can handle.

REFERENCES

[1] Sinha, R. M. K., & Thakur, A., 2005c, Divergence patterns in machine translation between Hindi and English, Proceeding of MT Summit X. Phuket, Thailand, pp. 346-353

[2] S. B. Kulkarni, P. D. Deshmukh, M. M. Kazi, K. V. Kale, “Linguistic to Socio-And-Psyco Linguistic Aspects in English-To-Marathi Language Translation”, International Journal of Research in Computer Applications And Robotics, 2013; 1(9), pp.197-205

[3] S. B. Kulkarni, P. D. Deshmukh and K. V. Kale, “Syntactic and Structural Divergence in English-to-Marathi Machine Translation”, IEEE 2013 International Symposium on Computational and Business Intelligence, August 24-26, 2013, New Delhi, pp. 191-194,doi: 10.1109/ISCBI.2013.46

[4] G.V. Garje, G.K. Kharate,”Challenges in Rule Based Machine Translation from English to Marathi”, 3rd International Conference on Recent Trends in Engineering &Technology (ICRTET’2014),pp. 243-248.

[5] Namrata G Kharate1 ,Dr.Varsha H. Patil2 “Survey of Machine Translation for Indian Languages to English and Its Approaches” International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,Volume 3,Issue 1,ISSN : 2456-3307,pp. 613-622.

[6] Joshi A., Sasikumar N. Constructive approach to teach inflections in Marathi language, Proceedings of National Conference on Advances in Technology andRecent Developments, Mumbai, India, 2008, pp.10-16

[7] Khan Md., Anwarus S., Amada S., Nishino T. Sublexical Translations for low-resource language, Proceedings of Workshop on Machine Translation andParsing in Indian Languages (MTPIL-2012), 24th International Conference on Computer Linguistics (Coling12)

[8] M. R. Walimbe. Sugam Marathi VyakranLekhan, G.Y. Rane Publication

[9] Wren P., Martin H. High School English Grammar and Composition, S Chand Publication

[10] CharugatraTidke, Shital B, Shivani P (2013) "Inflection Rules for English to Marathi Machine Translation"IJCSMC, Vol. 2, Issue. 4, April 2013, pg.7 – 18

[11] EshaPalta IITB. Word Sense Disambiguation, 2006-07, Master of Technology First Stage Report.

[12] Walker D. and Amsler R. 1986. The Use of Machine Readable Dictionaries in Sublanguage Analysis. In Analyzing Language in Restricted Domains, Grishmanand Kittredge (eds), LEA Press, pp. 69-83

AUTHORS

Ms.NamrataKharate is research scholar atdepartment of Computer Engineering, MCOERC, Nashik affiliated to SavitribaiPhule Pune University. Her topic of research is in Natural Language Processing domain. She is currently working asAssistant Professorat Vishwakarma Institute of Information and Technology, Pune

Professor Dr. V. H. Patil is currently Chairman Board of Studies (Computer Engineering) SavitribaiPhule Pune University. At SPPU, she is also member of Academic Council, Faculty and Research & Recognition committee. Having 29 years of teaching and academic administration experience, she is currently working as Professor & Head of Computer Engineering department with additional responsibility of vice-principal at Matoshri College of Engineering & Research Centre, Nashik(MS, India). She is recipient of various honours & awards. She has authored books in areas Discrete Mathematics (McGraw Hill) and, Data Structures using C++(Oxford University Press).