International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 10, Issue 9, September 2021 ISSN 2319 - 4847

Analysis of Indian Languages for Multilingual Machine Translation

Madhura Phadke[1], Satish Devane[2]

[1] Mumbai University, DMCE, Sector-3, Airoli, Navi Mumbai-400708

[2] Mumbai University, DMCE, Sector-3, Airoli, Navi Mumbai-400708

ABSTRACT

The Indian linguistic landscape gives broad perspective of variety of languages used by people around .Hindi is recognized as national language and English identified at national level as subsidiary official language. Most of the states make the language spoken by most of its people a official state language. Like Marathi in Maharashtra, kannada in Karnataka n so on. Thus unlike most of monolingual countries there is no single language in India. This possess a unique challenge for language processing mainly due to diversity in languages used. In this paper we focus on different language pairs from translation perspective. The diversity in language structure, vocabulary, syntactic and semantic variances demands increased efforts to deal with translation task.

Keywords: language, machine translation, monolingual, semantic variances

1. INTRODUCTION

India's society, culture, history and politics have continuously been shaped by the multiplicity of her languages. The country is home to speakers of about 461 languages. Of these, 447 languages are actively used in daily communication, while 14 are extinct - they no longer fulfil any communication need. Among these, 121 languages have more than 10,000 speakers and 22 of these are officially recognised in the Indian Constitution [1]. These include Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Maithili, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu. , languages can be classified into various 'families' based on the genealogical similarities among them. The main language families of India are the following: Indo-Aryan - this family includes major languages such as Hindi, Punjabi, Nepali, Marathi, Oriya, Bangla and Axomiya as well as tribal languages such as Bhili and Katkari. The Dravidian family of languages includes four major, literary languages in southern India – Tamil, Malayalam, Kannada and Telugu - as well as a number of tribal languages such as Toda in the Nilgiri Hills and Gondi in central India. The Daic family of languages in Arunachal Pradesh and in Assam and the Andamanese language family in the Andaman Islands are two smaller genealogical groups in the country. Not all languages are written, some are used only for verbal communication. The efforts by the government to bring the tribal communities in mainstream such as adivasi needs to overcome the language barrier. Translation helps in such scenarios.

2. LITERATURE REVIEW

In a multilingual society, information sharing takes place with variety of languages. Linguistics is the scientific study of language in all its facets. Language is a fundamentally important aspect of human life, and impinges on virtually everything that we do. Thus, Linguistics is a study which shares interests with a very wide range of other disciplines, and usefully complements a variety of other subject areas, such as the language subjects, Philosophy, Education, Sociology, Social Anthropology, Psychology and Artificial Intelligence.

To have the idea of existing systems being used for translation, we undergo a detailed survey of existing

systems.[2][3][4] The findings of the literature review are mentioned in table 2.1.

	Translation System	Year	Authors	Source Language	Target Language	Details
A)	**Direct Machine Translation Systems**					
1.	Anusaaraka systems among Indian Languages	1995	Rajeev Sangal	Telugu, Kannada, Bengali, Punjabi and Marathi	Hindi	The output of the system followed the grammar of the source language only. Developed by IIT Kanpur (earlier),IIIT Hyderabad(Now)
2.	Punjabi to Hindi MT System	2008	G S Josan and G S Lehal	Punjabi	Hindi	Based on direct word-to-word MT approach. Accuracy of this system is 90.67%. Developed by Punjabi University, Patiala.
3.	Web based Hindi-to-Punjabi MT System	2010	Goyal V and Lehal G S	Hindi	Punjabi	Extended version of Hindi-to-Punjabi MT System to Web. Developed by Punjabi University, Patiala.
4.	Hindi-to-Punjabi MT System	2011	Goyal V and Lehal G S	Hindi	Punjabi	The translation accuracy of the system is 87.60% Developed by Punjabi University, Patiala.
B)	**Transfer-Based MT Systems**					
1.	Mantra MT	1997	Bharati	English	Hindi	Uses XTAG based super tagger and light dependency analyzer for performing analysis of the input English text.
2.	MANTRA MT	1999	Hemant Darbari and Mahendra Kumar Pandey	English	Hindi, Bengali, Telugu, Gujarati	Translates in specific domain of personal administration that includes gazette notifications, office orders, office memorandums and circulars Uses TAG and LTAG to represent English & Hindi grammar. It is based on synchronous Tree Adjoining Grammar and uses tree transfer for translating from English to Hindi.
3.	An English–Hindi Translation System	2002	Gore L and Patil N	English	English	Uses different grammatical rules of source and target languages and a bilingual dictionary for translation. The domain of the system was weather narration
4.	MAT	2002	Murthy K	English	Kannada	Uses UCSG(Universal Clause Structure Grammar), morphological analyzer & post-editing

5.	Shakti	2003	Bharati, R Moona, P Reddy, B Sankar, D M Sharma and R Sangal	English	Indian languages	Combines linguistic rule-based approach with statistical approach. The system consists of 69 modules
6.	English-Telugu MT System	2004	Bandyopadhyay S	English	Telugu	Uses dictionary containing 42,000 words. A word form synthesizer for Telugu is developed and incorporated in the system.
7.	Telugu-Tamil MT System	2004	Bandyopadhyay S	Telugu	Tamil	Uses the Telugu Morphological analyzer and Tamil generator for translation. The system makes use of Telugu-Tamil dictionary. It also uses verb sense disambiguation.
8.	OMTrans	2004	Mohanty S, Balabantaray R C	English	Oriya	Based on grammar and semantics of the source and target language. Uses WSD too.
9.	The MaTra System	2004, 2006	Ananthakrishnan R, Kavitha M, Hegde J J, Chandra Shekhar, Ritesh Shah, Sawani Bade, and Sasikumar M	English	Hindi, Bengali, Telugu, Gujarati	The domain of the system is news, annual reports and technical phrases It has different dictionaries for different domains. Requires considerable human assistance in analyzing the input. Uses sentence splitter.
10.	English-Kannada machine-aided translation system	2009	K Narayana Murthy	English	Kannada	The domain is of government circulars. Uses Universal Clause Structure Grammar (UCSG) formalism. The system is funded by the Karnataka government
11.	Tamil-Hindi Machine-Aided Translation system	2009	Sobha L, Pralayankar P and Kavitha V, Prof. C N Krishnan	Tamil	Hindi	Based on Anusaaraka. Uses a lexical-level translation and has 80-85% coverage

12.	Sampark System: Automated Translation among Indian Languages	2009	Technology for Indian Languages (TDIL) project	English	Indian Languages	Uses Computational Paninian Grammar (CPG) for analyzing language and combines it with machine learning. It is developed using both traditional rules-based and dictionary-based algorithms with statistical machine learning.
C)	**Interlingua Machine Translation Systems**					
1.	ANGLABHARTI	2001	R M K Sinha, Jain R, Jain A	English	Indian Languages	Developed using pseudo-interlingua approach. The domain of this system is public health
2.	UNL-based English-Hindi MT System	2001	Dave S, Parikh J and Bhattacharyya P	English, Hindi	Hindi, Bengali, Marathi	Uses Universal Networking Language (UNL) as the Interlingua structure. Developed by IIT Mumbai.
3.	AnglaHindi	2003	R M K Sinha and Jain A	English	Indian Languages	Pseudo interlingual rule-based English to Hindi Machine-Aided Translation System.
D)	**Hybrid Machine Translation Systems**					
1.	ANUBHARATI Technology	1995, 2004	Sinha	Hindi	Indian Languages	A combination of example-based, corpus-based approaches and some elementary grammatical analysis
2.	ANUBHARTI-II	2004	R M K Sinha	Hindi	Indian Languages	Uses Generalized Example-Base (GEB) along with Raw Example-Base (REB) MT approach for hybridization
3.	Bengali to Hindi MT System	2009	Chatterji S, Roy D, Sarkar S and Basu A	Bengali	Hindi	Uses an integration of SMT with a lexical transfer based system (RBMT)
4.	Lattice Based Lexical Transfer in Bengali Hindi MT Framework	2011	Sanjay Chatterji, Praveen Sonare, Sudeshna Sarkar, and Anupam Basu	Bengali	Hindi	Uses transfer based MT approach with the help of lattice-based data structure
5.	A web based English to Punjabi MT system for News Headlines	2013	Harjinder Kaur, Dr. Vijay Laxmi	English	Punjabi	Using rule based approach the system parses the source text and produces as intermediate representation.
6.	Transmuter : An approach to rule based English Marathi Translation system	2014	G. Garje	English	Marathi	Focus is on grammar structure of target language that produces better and smoother translation.

E)	**Example Based Machine Translation (EBMT) Systems**					
1.	ANUBAAD	2000, 2004	Bandyopadhyay S	English	Bengali	Domain specific to English Headlines translation Example-base, Generalized Tagged example-base and Phrasal example-base are separately maintained. If the headline cannot be translated using above methods then the heuristic translation strategy is used
2.	VAASAANUBAADA	2002	Vijayanand K, Choudhury S I and Ratna P	Bengali	Assamese	Domain limited to News Text Sentence level Machine Translation for Bengali Includes pre-processing and post-processing tasks. Uses bilingual aligned corpus
3.	Shiva and Shakti MT System	2003	CMU USA, IIIT Hyderabad and IISC Bangalore, India	English	Hindi, Marathi and Telugu	Uses combination of Example-based, rule based and statistical approaches.
4.	ANGLABHARTI-II	2004	R M K Sinha	English	Indian languages	Uses Generalized example-base (GEB) approach and Raw Example-Base (REB) Contains the modules for an error analysis and post-editing automatically.
5.	Hinglish machine translation system	2004	Sinha and Thakur	Hindi	English	Based on AnubBarti-II and AnglaBharti-II Performs very shallow grammatical analysis
6.	English to {Hindi, Kannada, Tamil} and Kannada to Tamil Language-Pair	2006	Balajapally P,P Pydimarri, M Ganapathiraju, N Balakrishnan and R Reddy	English Kannada	Hindi, Kannada and Tamil Tamil	Based on a bilingual dictionary comprising of sentence dictionary, phrases dictionary, words dictionary and phonetic dictionary.
7.	The MATREX System	2008	Ankit Kumar Srivastava, Rejwanul Haque, Sudip Kumar Naskar and Andy Way	English	Hindi	Uses marker based chunking and "edit-distance style" dynamic programming alignment algorithm Domain limited to Conference papers
F)	**Statistical Machine Translation Systems**					
1.	Shakti	2003	Bharati, R Moona, P Reddy, B Sankar, D M Sharma and R Sangal	English	Indian Languages	Combines linguistic rule based approach with statistical approach
2.	English to Indian Languages Machine Translation System	2006	Consortium of Nine institutions	English	Indian Languages	Limited to Tourism and Healthcare domain Uses statistical techniques and tools including the POS tagger, parser , decoder

3.	English to Malyalam Translation	2008	Mary Priya Sebastian, Sheena Kurian K, G. Santhosh Kumar	English	Malyalam	Monolingual corpus of Malyalam is used and bilingual is used for English. Order conversation rules are used to overcome the structural differences.
4.	Punjabi to English Machine Transliteration system for proper noun	2013	Pankaj Kumar, Vinod kumar	Punjabi	English	The system is divided into two parts learning and transliteration.
5.	Google Translate	2006-2021	Google	Multiple Languages	Multiple Languages	Earlier was developed as SMT, recently many languages are added with the support of Neural network.

Table 2.1 Machine Translation Systems developed in India

3. STUDY OF LANGUAGES FOR EFFECTIVE TRANSLATION

To begin with the development, we need to study the formation of sentences in each of the languages. There are 8 part of speech components (POS) namely ([noun, pronoun, adjective, verb, conjunction, preposition, interjection]). So each word in a sentence falls under one of this category. The important part in translation is that a word may change its role from one POS to another, when the sentence is translated from one language to another. [5][6] Hence POS handling plays important role in MT.

This and many other issues we came across in the current development stage of our MT system. To summarize, we have listed it.

1. Divergences between different languages.
2. Lexico Semantic divergence
3. Structural and Syntactic Divergence
4. Commonly Found Divergence

3.1 Word ordering in linguistics

The syntactic structure of a language is determined by the word order. Words are classified into 8 parts-ofspeech (POS) [noun, pronoun, adjective, verb, conjunction, preposition, interjection]. [7][8] The arrangement of these POS in sentences is determined according to the structure the language follows. English follows Subject-VerbObject (SVO) structure while Marathi follows Subject-Object-Verb (SOV) structure [2].Along with Marathi other Indo-Aryan language like Hindi, also follow the SOV structure.

3.2 Importance of adpositions in linguistics

Adpositions are words which can occur before or after a phrase, word, or a clause that is necessary to complete the meaning of a given sentence. Adpositions are mainly categorized as:

- Prepositions
- Postpositions
- Circumpositions

3.3 Prepositions

Prepositions are defined as the words placed before the complement. Prepositions are used in English.

Example: I value my family above everything else.

a. Postpositions:
Postpositions are words which come after the complement. Postpositions are used in Marathi, Hindi, Urdu, Korean, Turkish, and Japanese.

b. Circumpositions:
Circumpositions are words that appear on both sides of the complement. They are used in English, Dutch, Swedish, and French.

Example: I will exercise regularly from now on.

The languages which follow SOV Structure use postpositions. Hence, while translating an English sentence.[9][10] (SVO structure) to a Marathi sentence (SOV structure), we need to change the prepositions (of English) to postpositions (of Marathi). This is a major issue which needs to be resolved for inflecting the nouns, verbs and cases (Vibhakti).

- Inflection

Generating inflection of a word is important to retain the correct form of the word in Marathi.[11] Words can be classified in two types depending on the Inflection as :
Inflectional Words: • Noun • Pronoun • Adjective • Verb

- Non-Inflectional Words: • Adverb • Preposition • Interjection • Conjunction

The words are inflected on the basis of changing Gender (Masculine, Feminine, Neuter), Multiplicity (Singular, Plural), Tense (Present, Past, Future), and Case (Nominative, Accusative, Instrumental, Dative, Ablative, Genitive, Locative, Vocative).

Languages are bound to syntax, word structure that is morphology, sound structure known as phonology and vocabulary lexicon [12][13]. Translation is greatly affected the way all these aspects are looked into. Based on the language family, structure of the sentence varies. Position of the word may change depending upon the choice of source and target language being considered. Some words in a languages represents several contexts in other [14][15]. Handling all these in an appropriate way improves the translation quality. The challenges to accommodate this in translation process are discussed below with some principle features of different languages to demonstrate the complexity.
Translation tools fails to support all the flavors of underlined language, due to the vast features that need to work upon. For example consider the example shown in Figure 2.1 and Figure 2.2 for Google Translate between Marathi and Hindi. In the first example Marathi is flexible for it's structure. It supports Subject – Object – Verb, Subject- Verb- Object, Subject – Verb and several other sentence syntax. Though Hindi also supports variety of sentence structures, it fails in translation for many cases. Identifying the causes forms the basis for challenges in machine translation. In case of human expert translation the language fluency, knowledge of the same, background of the commination, context awareness helps the translator to produce the appropriate translation. Many of these factors are not adopted by the system used for translation.

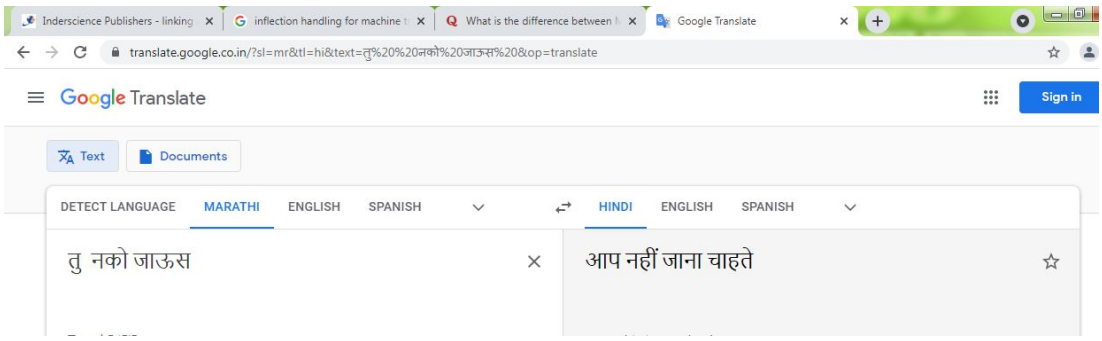

Figure 3.1 Google Translate Marathi-Hindi (sentence structure)

In example demonstrated in 2.1, the message to be communicated is "You don't go" as stated in source language

Marathi. Due to the underlying translation system it is translated in a way which states "You are not willing to go." The translation in this case fails to maintain the meaning of said thing. And the consequences maybe harmful in many cases if the intended thing is not communicated and miss-interpreted by the receiver.

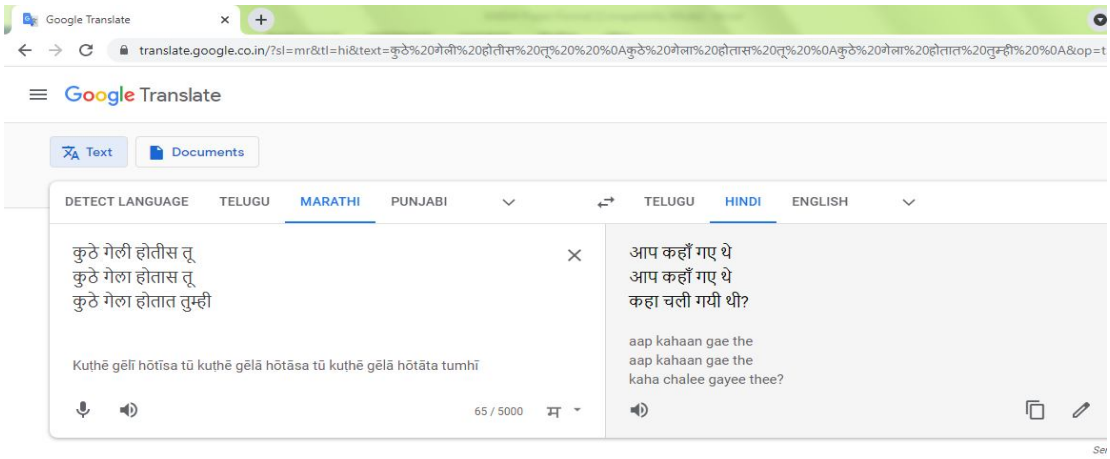

Figure 3.2 Google Translate Marathi-Hindi (verb-pronoun relation)

Figure 2.2 demonstrates another case where the translation fails to identify the gender of the subject involved in context. In the first sentence it's about the female whereas next sentence talks about male gender. But as can be seen in translation both are translated in same manner neglecting the important information present in context. The third case describes plural case, gender information is not clearly mentioned that needs to be retrieved form context, is translated by assuming the female gender subject.
Many such language issues needs to work on to have the meaningful translation. Here we present the divergences for Marathi, Hindi , Bangla and English.

Language pair	Example
Marathi - Hindi	उडत आम
Hindi - Marathi	जब हम लऱए यह पडत .
Marathi - English	I would like a mango

	Someone killed Rama
English - Marathi	switch off button ऑफ बटण I feel awkward

Table 2.2 Challenging translation examples

During our study we have tested the existing system for variety of sentence types. In table 2.2 we show few examples to have the glimpse of translation challenges. The major challenges are as follows :

1. Out of dictionary word. All the words in source language are not included in corpus of target language which leads to wrong translation.
2. Omitting words during translation, which plays important role in representing the context.
3. Transliterate instead of translate. Here the sentence is transliterate using the pronunciation.
4. Syntactic divergence
5. Morphological divergence etc.

4. DISCUSSIONS AND FINDINGS

Machine translation (MT) is still a huge challenge for both IT developers and users. From the beginning of machine translation, problems at the syntactic and semantic levels have been faced. Today despite progress in the development of MT, its systems still fail to recognize which synonym, collocation or word meaning should be used.. The paper deals with the analysis of machine translation of general everyday conversation for three languages. The results of the analysis show that incorrect translation is still the issue with translation system.

5. CONCLUSION

The work in translation era is challenging. We are working on methodology which helps improving the performance of translation task. As discussed in this paper several issues needs to be worked upon to achieve the required quality of translation. Languages under study plays major role in developing a system. Language diversity needs to be studied in depth to have the system working as per expectations.

References

[1] Nida, Eugene & Charles R. Taber. “The Theory and Practice of Translation. Leiden: Brill”, 1982.
[2] Ana Fernández Guerra “Translating culture: problems, strategies and practical realities”, a journal of literature, culture and literary translation No. 1 - Year 3 12/2012 - LT.1
[3] Vishal Goyal & Gurpreet Singh Lehal, (2009) “Advances in Machine Translation Systems”, National Open Access Journal, Volume 9, ISSN 1930-2940 http://www.languageinindia
[4] Sanjay Kumar Dwivedi & Pramod Premdas Sukhadeve, (2010) “Machine Translation System in Indian Perspectives”, Journal of Computer Science 6 (10): 1082-1087, ISSN 1549-3636, © 2010 Science
[5] Aditi Kalyani & Priti S., " A Review of Machine Translation Systems in India and different Translation Evaluation Methodologies", International Journal of Computer Applications (0975-8887) volume- 121. No. 23, July 2015
[6] Amruta Godase & Sharvari Govilkar “Survey Of Machine Translation Development For Indian Regional Languages” IJMTER-2014

[7] G.V.Garje, G.K.Kharate & H.Kulkarni "Transmuter: An Approach to Rule-Based English to Marathi Machine Translation" IJCA (0975-8887) 2014
[8] Charugatra Tidke, Shital Binayakya, Shivani Patil , Rekha Sugandhi "Inflection Rules for English to Marathi Translation" IJCSMC, Vol. 2, Issue. 4, April 2013, pg.7 – 18
[9] Cheragui M.A., "Theoretical Overview of Machine Translation", Proceedings ICWIT, 2012.
[10] John Hutchins, MT News International, no. 14, June 1996, pp. 9-12
[11] The encyclopedia of languages and linguistics, ed. R.E.Asher (Oxford: Pergamon Press,1994), vol. 5, pp. 2322-2332
[12] John Hutchins , Practical experience of machine translation: proceedings of a conference, London, 5-6 , November 1981;
[13] https://translate.google.co.in/#en/mr/monkey%20ate%20banana%20as%20it%20was%20ripe
[14] Hutchins J.W., "Introduction to Machine Translation", Academic Press, 1992.
[15] http://language.worldofcomputing.net/machine-translation/challenges-in-machine-translation.html

AUTHOR

Profesor Madhura Phadke is pursuing Ph.D. She has 21 years of teaching experience. She is good in various subjects such as machine learning, security and database. Her research is mainly focused on machine translation using machine learning. She has published 11 papers in international conference, 4 international journal, 13 in national conferences.

Professor Dr. S. R. Devane is an Academician of the IIT (Ph. D: Information Technology | M.E: Electronics | B.E.: Elcctronics |) and principal of KBTCOE, Nashik. Professor Devane is proficient in many technical areas such as networking, Artificial Intelligence, Data Mining etc. He has published 12 papers in international