Turkish Journal of Computer and Mathematics Education Vol.12 No. 5 (2021), 1134-1147

Research Article

A Review of Machine Translation for South Asian Low Resource Languages

Syed Abdul Basit Andrabi[a*], Abdul Wahid[b]

[a*]Department of CS&IT, Maulana Azad National Urdu University, Hyderabad Telangana, India.
E-mail: sbasit.11@gmail.com
[b]Department of CS&IT, Maulana Azad National Urdu University, Hyderabad Telangana, India.
E-mail: wahidabdul76@yahoo.com

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Machine translation is an application of natural language processing. Humans use native languages to communicate with one another, whereas programming languages communicate between humans and computers. NLP is the field that involves a broad set of techniques for analysis, manipulation and automatic generation of human languages or natural languages with the help of computers. It is essential to provide access to information to people for their development in the present information age. It is necessary to put equal emphasis on removing the barrier of language between different divisions of society. The area of NLP strives to fill this gap of the language barrier by applying machine translation. One natural language is transformed into another natural language with the aid of computers. The first few years of this area were dedicated to the development of rule-based systems. Still, later on, due to the increase in computational power, there was a transition towards statistical machine translation. The motive of machine translation is that the meaning of the translated text should be preserved during translation. This research paper aims to analyse the machine translation approaches used for resource-poor languages and determine the needs and challenges the researchers face. This paper also reviews the machine translation systems that are available for poor research languages.

Keywords: SMT, RBMT, Natural Language Processing, Neural machine translation. Low Resource Languages.

1. Introduction

Machine translation is the technique of translating the text of one natural language into another natural language by using computer software, e.g. English to Urdu. It is an automated process in which the computer does the translation work. Machine translation is an application of computer linguistics [1]. Computer linguistics is an interdisciplinary field that requires language and computer experts. Language experts frame the rules of the languages and computer experts program the computer to understand these rules. The area of machine translation started when electronic computers came into existence. The concept of machine translation was first used during World War II by Weaver, one of the pioneers in machine translation to crack the German enigma code. In the 1950s, the field of machine translation became a reality with the demonstration of Georgetown Experiment, which translated more than sixty Russian sentences into English automatically [2]. As a result, a lot of interest and funding flowed in for almost a decade. The United States was the leading research and funding agency with the primary aim to strengthen their military and defence intelligence. But the research in machine translation came to a halt for about decade in 1966 after the (ALPAC) "Automatic Language Processing Advisory Committee" report.

According to the ALPAC report, machine-translation output was costly, and output was not faster than full human translation because, in machine translation, there was a post-editing requirement. Hence, there was no advantage in using machine translation and suggested that funding should go to basic linguistic research to improve human translation compared to Machine human translation. Due to the recent industrial growth, there is a significant impact on machine translation. The need arises that requires contents to be available in all regional languages worldwide [3]. The beginning years of research in this field were dedicated towards the rule-based systems. During the 1980s due to increasing computational power, there was a transition from the rule-based system to statistical machine translation approach. The enormous increase in the electronic text's multilingual data has ignited plenty of monolingual and cross-lingual information retrieval efforts. It is vital to share information with people for their development [3]. Suppose the MT researchers can develop a multilingual machine translation model. In that case, individuals with various dialects can share their insight and thoughts worldwide in their local dialect. Everybody in the globe can have the ability to get this information and ideas in their local dialect. The translation process's purpose is that meaning of the translated text should be same as that of the original. One advantage of translation is the accessibility of information in the birth languages. Due to technology limitations, it has not been possible to generate information in many languages of the world. The majority of the research in the last few decades was dedicated towards the automatic (NLP) natural language processing for English, East Asian and European languages. Still, unfortunately, South Asian languages received less attention [4]. Due to digital resource scarcity, machine translation is a challenging task for resource-poor languages.

2. Machine Translation Approaches

The Machine translation approaches are classified as Rule-Based Machine Translation" (RBMT), corpus-based, Hybrid and knowledge-based approach [5] [6]. The classification is shown in Figure 1.

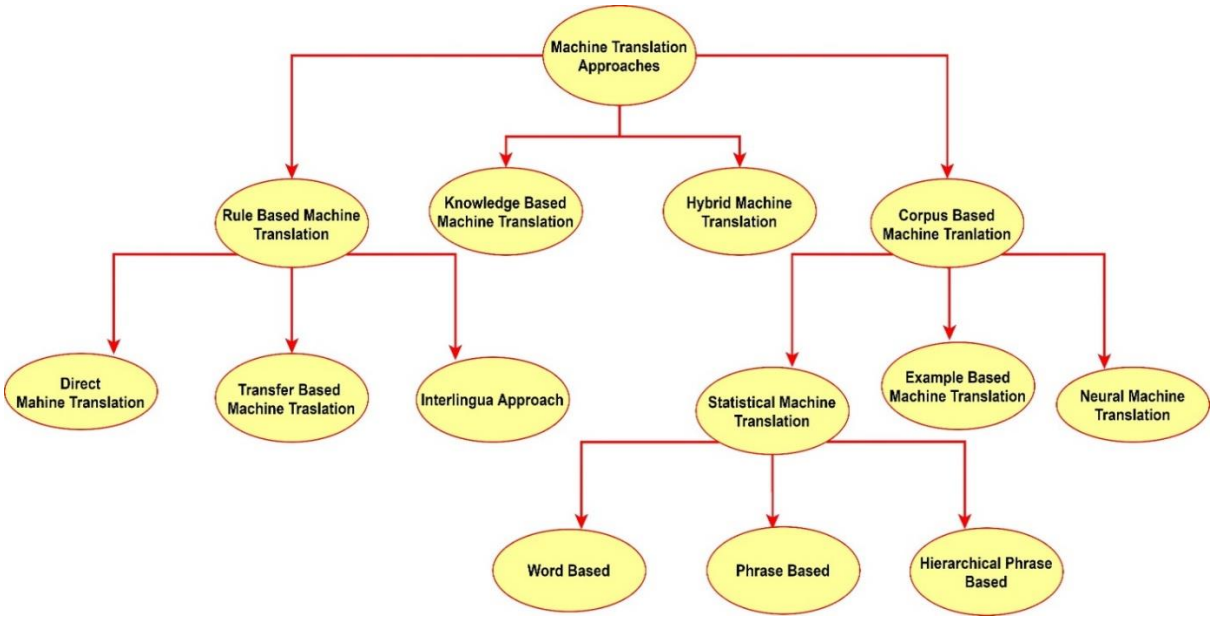

Figure 1: Approaches to Machine Translation

3. Classification of Machine Translation approaches based on RMBT

The rule-based machine translation techniques can be further divided into three categories based on Bernards pyramid, as shown in Figure 2.

3.1 Direct Machine Translation

This Machine Translation approach (MT) operates at the lowest level of the machine translation pyramid given by Bernard Vauqous, as shown in figure 2. It is one of the oldest methods that work at the word level and uses a bilingual dictionary to directly mapource language and target language [7]. This approach does not perform structural and morphological analysis of source language. Hence this approach does not give good results [8].

3.2 Transfer Approach

This Machine Translation approach operates at level 2 of the machine translation pyramid in which source language is first converted into an intermediate language. Then it is used to generate target text using a bilingual dictionary. This approach works in three phases: Analysis, Transfer and Generation. The source language text is first analysed using linguistic information to form a syntactic representation of source language with source language parser. After the intermediate representation, the next stage is to convert it into the target language representation, where the transfer stage comes into play. Previous source syntactic representation is converted into target syntactic representation. The last step is the generation stage, where the target-language text is generated with morphological analysis. [5]

3.3 Interlingua Approach

The Interlingua approach is similar to the transfer based approach, but in this case, an extensive syntactic, semantic and morphological analysis of source language is done. In this case, the text to be translated is converted into an intermediate form called meta-language or an Interlingual language, which is language-neutral representation. The next step is to generate the target language from the intermediate representation. In this case, a thorough analysis of source language is done [9].

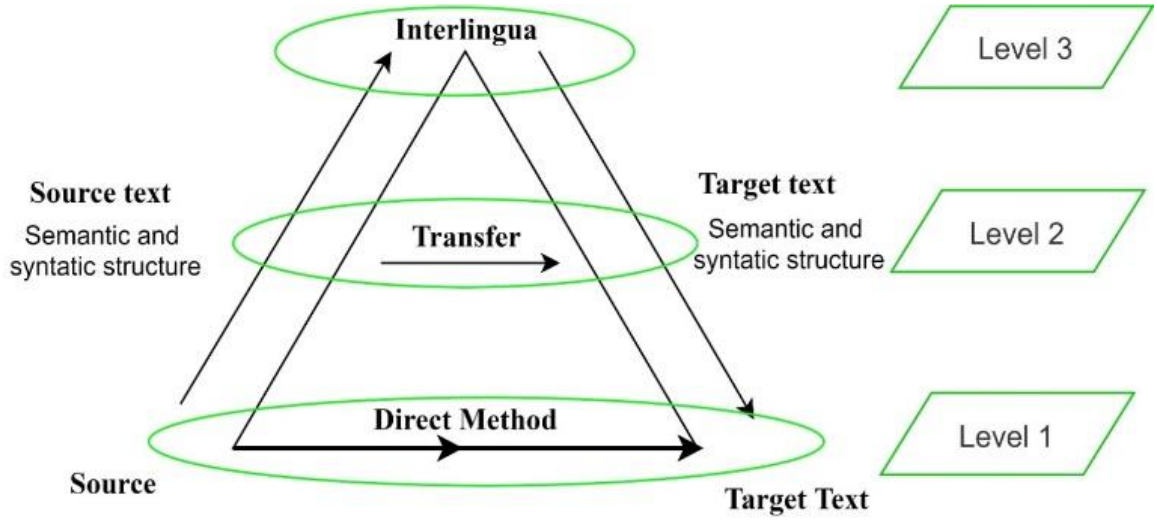

Figure 2: Machine Translation Paradigm

4. Knowledge Base Machine Translation

Knowledge Base Machine Translation consists of a huge knowledge base containing parallel sentences and an inference engine. The problem with this type of approach is that it is difficult to represent knowledge and define its granularity.

5. Hybrid Approach

The hybrid approach uses two or more machine translation methods like SMT and RBMT or RBMT and EBMT. The accuracy of the hybrid approach is reasonable compared to other methods, but it is costly during the initial stage.

6. Corpus-Based Machine Translation (CBMT)

The CBMT known by the name of data-driven approach or empirical machine translation. This approach overcomes specific problems of machine translation approaches that were based on Bernards pyramid. In this approach, there is no need for syntactic, semantic and morphological analysis. In this case, a huge amount of corpus is required for good quality output. Furthermore, the corpus-based machine translation can be divided into three different types which are as follows:

6.1 Example-Based Machine Translation

The example-based machine translation is a subtype of corpus-based machine translation and does not require any dictionary and grammatical rules. This type of machine translation is based on the database approach, where we have many examples stored that are already translated. If a new sentence is encountered, such past translations are used, and the best matching algorithm is applied to get the translation of the new sentence.

6.2 Statistical Machine Translation

This is another machine translation approach that comes under the corpus-based machine translation approach. In this approach, we require a huge bilingual corpus to train the system, and no rules and Grammar are needed in this approach. This model learns mappings from the parallel corpus and then uses these learned mapping to translate new sentences. The SMT consists of three main components: language model, translation model, and decoder [13]. The increase in corpus size in SMT increases the BLEU score as the corpus size has a significant impact on the BLEU score.

6.3 Neural Machine Translation (NMT)

Neural Machine Translation is a promising approach that uses artificial neural networks and substantial parallel corpus. The excellence of neural machine translation is that it is based on end to end learning. The NMT makes use of encoder-decoder architecture.

Figure 3 shows the evolution of machine translation approaches. From 2014 onwards, two promising approaches are SMT and NMT Nowadays, the two main approaches of machine translation used are Statistical Machine Translation and Neural Machine Translation. In this paper, we elaborate on these two approaches only.

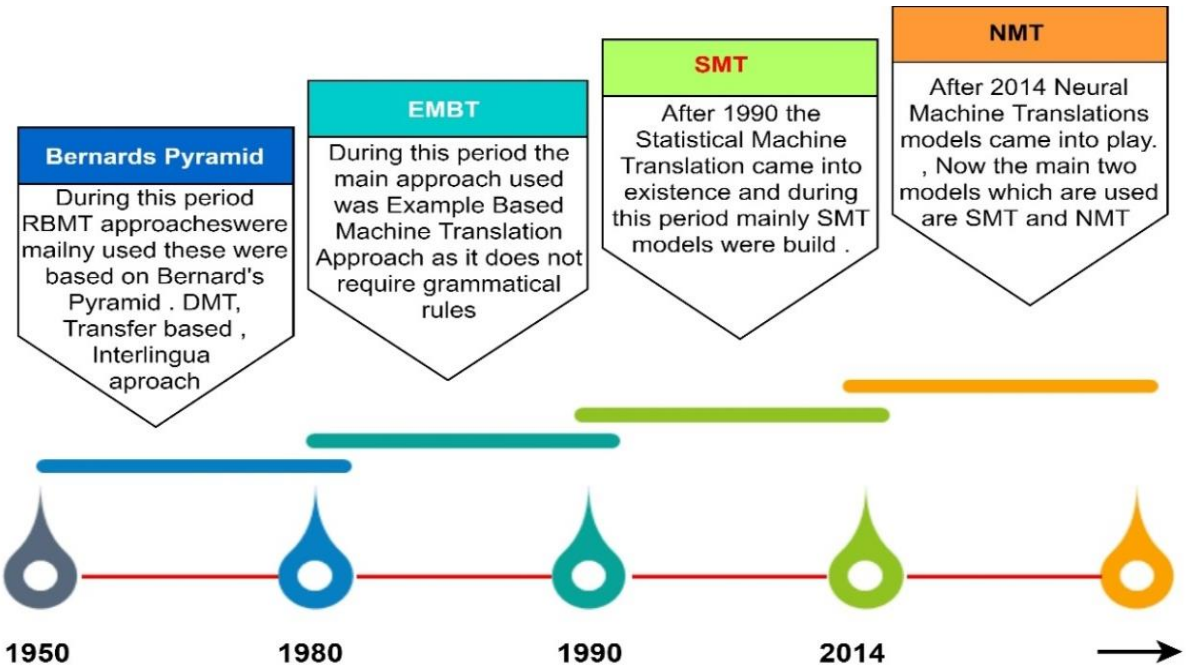

Figure 3. Evolution of Machine Translation Approaches

7. Statistical Machine Translation Approach (SMT)

This is the data-driven approach based on the Statistical Models and Noisy Channel model for communication, which was introduced by Shanon in 1948. SMT is based on Bayes Theorem of Probability which uses an enormous Bilingual corpus to derive the rules and mapping of Source language and Target Language. MT is still a promising approach because of its several advantages like low cost, rapid prototyping, uses human translation as its building block and also supports many languages which do not have enough lexical resources. This approach produces the best results when large datasets are available.

From the concept of Shanon's noisy channel model, consider a distorted message R (Foreign String f) a model to know how the message is distorted (translation model t (f|e)) and also a model on which original messages are probable (language model p(e)). Our objective is to retrieve the original message S (English string e), shown in figure 4 below.

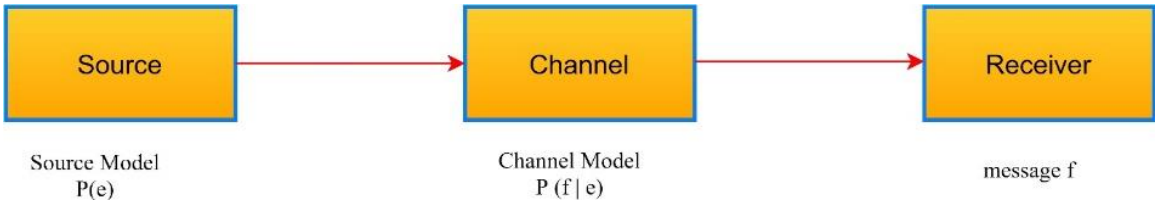

Figure 4. Machine translation as Noisy channel model.

From the probabilistic view, the model can be represented as follows:

Given an English sentence e, we seek the foreign sentence f that maximises P (f | e) which is the most likely translation we can write as $\underset{f}{\operatorname{argmax}} p(f \mid e)$ which means the foreign sentence f, out of all sentences which produces the highest value for P(f | e)

$$P(f \mid e) = \frac{P(f) * P(e \mid f)}{P(e)}$$

$P(f \mid e)$ is Posterior probability, $P(e \mid f)$ is a likelihood, $P(f)$ is prior probability, and $P(e)$ is a marginal probability

$$\underset{f}{\operatorname{argmax}}\, P(f|e) = \underset{f}{\operatorname{argmax}}\, P(f) * P(e|f) \quad (1)$$

The *P(e)* is not in the division because the most likely translation f maximise the product of two terms and will remain the same for all.

8. Neural Machine Translation (NMT)

Neural Machine Translation is a promising approach that maps words of source and target languages in the end to end fashion. It addresses the drawbacks of classical machine translation approaches. NMT architecture consists of encoder and decoder, two RNN (Recurrent Neural Networks), namely encoder and decoder. The encoder network takes input and creates a fixed-length vector, whereas the decoder generates translated output text from the encoded vector [16]. The architecture can be combined with the attention model to achieve excellent performance.

From the probabilistic viewpoint, translation is analogous to find a target sentence t that maximises the conditional probability of t for a given sentence s [17]i:e $\underset{t}{\operatorname{argmax}}$ P(t | s). The encoder reads source sentence S as a sequence of vectors S=(x_1, x_2, x_3……) in vector v, a standard RNN uses the following equation to compute the output.

$$h_t = s\,(W x_t + W h_{t-1}). \quad (2)$$

h_t is the hidden state at time t which is a nonlinear mathematical function of input x_t multiplied by weight matrix W added to the previously hidden state output (wh_{t-1})

RNN generalises feed forward neural networks that store previous input and combine it with the current input. In RNN, the Neural Network goes back and checks what has happened in the previous nodes before taking any decision.

8.1 Proposed Models of NMT

Several Architectures were proposed by the researchers in NMT some of them are mentioned in this paper. Bahdanau et al. proposed NMT by jointly learning to align and translate. This architecture belongs to the encoder-decoder model of RNN in which source sentence is encoded into a fixed-length vector from which decoder generates the target sentence. The problem in the encoder-decoder model of NMT is that it cannot handle long sentences. In this paper, authors have proposed a solution to align and translate mechanism jointly. But this approach will not work for the language whose secret is written in complex fashion as it is difficult to extract individual sequences of the language. This model was tested on English, and French languages, both Subject Verb Object (SVO) order and has not been used on different word order languages,hich is challenging in machine translation.

GNMT (Google's Neural Machine Translation): Google developed this Architecture in 2017 to bridge the gap between human translation and Machine Translation. This architecture consists of three components. The component is encoder, decoder, and attention network with 8 layers with LSTM (Long Short Term Memory Network) RNN units, which address the vanishing gradient in RNN. The attention network was added to the encoder-decoder model to increase the performance, and this architecture claims to achieve the accuracy rate of 60%. This model uses the sequence to sequence form of learning. It also worked at the symbolic level and focused on languages like French, Spanish, and Chinese.

Hierarchy to sequence Attention NMT Model: This model was proposed by Jinsong Su et al. in 2018. In this approach, the source sentence is divided into a sequence of different short clauses, which are translated sequentially. In this model, the bottom level RNN operates at the word level. The sentence S is divided into clauses c_1,c_2,c_3……c_n where each clause contains a sequence of words and at the end of each special clause token is placed to mark the end of the clause. At the decoder side, two attention networks can predict the next word based on previous given context and words generated previously. It chooses the clause length arbitrarily, and no mechanism is employed to detect optimal clause length. This model was also evaluated for Chinese-English and English-German languages.

9. Need for Machine Translation

The Internet World Stats Report describes that the content available on the internet in different languages varies, and the most dominant language on the internet is English [20], keeping in view this issue there is a dire

need of machine translation system to make the web content available to everyone in their native language. Below-given figure 5 shows the top ten languages on the internet in millions of users.

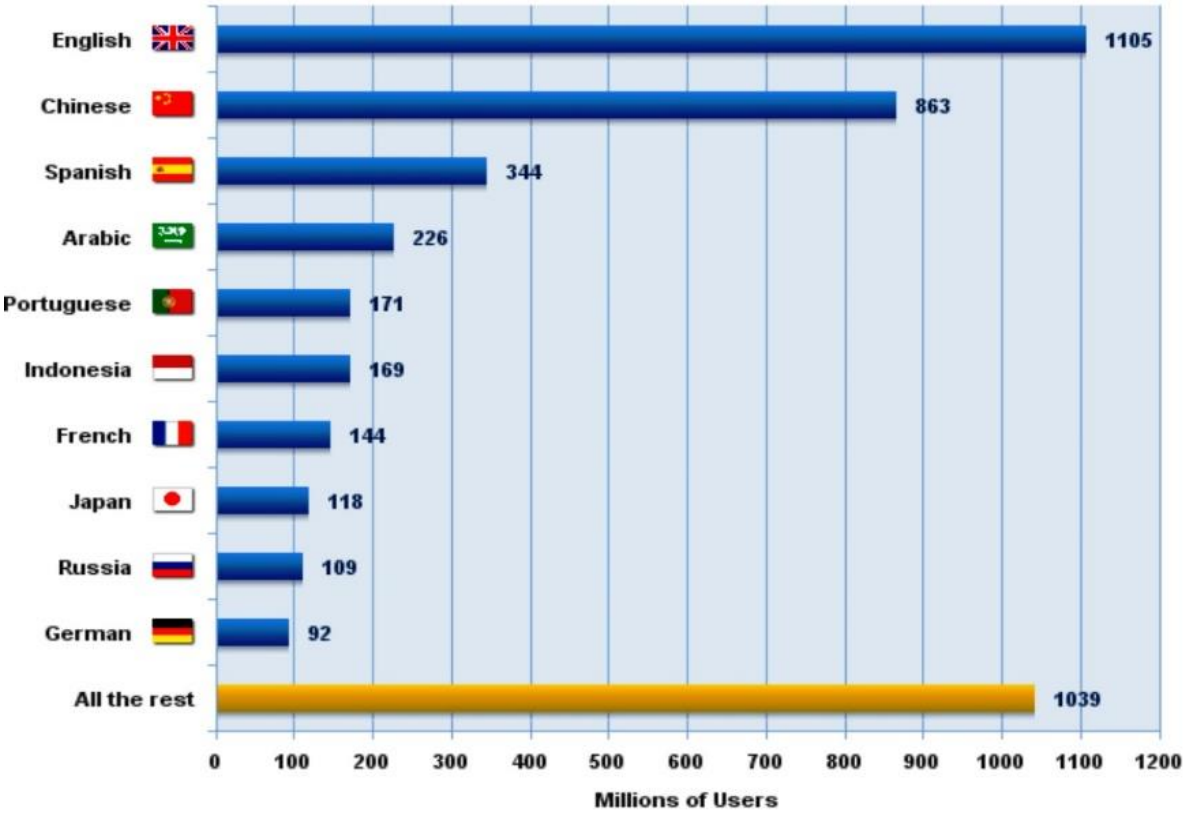

Figure 5. Languages used on the web [18]

Machine translation frameworks are expected to decode or translate creative works from any dialect to local dialect. Such machine interpretation frameworks can break the language obstruction by quickly making work accessible to the globe's masses. Numerous web pages may contain information related to our interest in a foreign dialect, and with the help of machine translation, we understand the content present in those web pages. Machine translation can also help commercial product manufacturers prepare product manual in many languages that an be used by different countries [13]. With the advancements in the internet, millions of users worldwide can get the information in their native language with the help of machine translation. In modern civilisation, machine translations have growing need and importance in economics, business and industrialisation. The social and political urgency of machine translation rises in societies where more than one language is spoken. [21]

In health care, machine translation plays a crucial role in upgrading access to multilingual health materials. Several machine translation systems are available, but their performance is not adequate in the public health domain [22]. During the last decade, machine translation technology has improved. Currently, machine translation is used by language providers and several companies and by the government departments.

Machine translation provides an economical means to translate an enormous amount of corpus from one language to another with less post-editing. It translates a vast amount of text in less time than a human translator, thus saving a lot of time.

10. Challenges in Machine Translation

Machine translation is a difficult and challenging problem. The difficulty of machine translation is to handle different ambiguities that are present in source and target languages. These ambiguities are either present naturally in sentences or arise due to the inability to form grammatical sentences. Natural languages have different aspects and feature i:e there is a difference in representing a concept in different languages. If one language represents a concept in one way, the other language may represent it differently. Some ambiguities that cause the problem in machine translation are below:

Lexical Ambiguity: Preferably, each word in language must have their unique meaning or sense; however, for natural languages, many words have multiple interpretations due to which sentence becomes unclear or vague. This type of ambiguity is lexical ambiguity. Lexical ambiguity can be of two types:

i. Word belongs to one or more lexical categories (noun, verb, adjective, etc.).
ii. One word has more than one understanding and belongs to the same lexical category. First, one can be solved by performing a syntactic analysis.

Table 1. Words belonging to a different lexical category with a different meaning

English Word	Lexical Category	Meaning	Meaning in Urdu
Light	Noun	This light is sufficient for reading	روشنی
Light	Verb	This shoe is light	ہلکا
Bear	Noun	Wild Animal	ریچھ
Bear	Verb	I cannot bear this	برداشت
Stick	Noun	The teacher beat him with a stick	چھڑی
Stick	Verb	Stick to your plan	قائم

Table 2. Words belonging to the same lexical category with a different meaning

English Word	Lexical Category	Meaning	Meaning in Urdu
Bat	Noun	Animal	چمگادڑ
Bat	Noun	Cricket bat	بلا
Nail	Noun	Iron Nail	کیل
Nail	Noun	Finger Nail	ناخن
Net	Noun	Volleyball net	جال
Net	Noun	Total	قل

Referential Ambiguity: This type of ambiguity deals with the use of the pronoun. For example, consider the sentence Raj went to Varun. He said I am tired in this sentence there is an ambiguity whether He is referring to Raj or Varun.

Word order issues: Word order issues are challenging for machine translation consider the two languages like English and Urdu, English follows (SVO) Subject verb object order whereas Urdu is free to order language and commonly used order is (SOV) Subject Object Verb, this also presents a challenge in Machine Translation. The example of the word order of English and Urdu is shown in Figure 6. Figure 6 shows that the English script is read and drafted from left to right, whereas script Urdu is read and drafted from right to left.

Figure 6. Word order issues in SVO and SOV languages

Prepositions and Post-Positions: Prepositions and post-positions also create complexity in machine translation. Some Languages use prepositions, and some use post-positions, the translation between these two languages is challenging. Consider the example of English and low resource Urdu language, as shown in Figure 7, the English language uses prepositions, and the Urdu language uses post-positions.

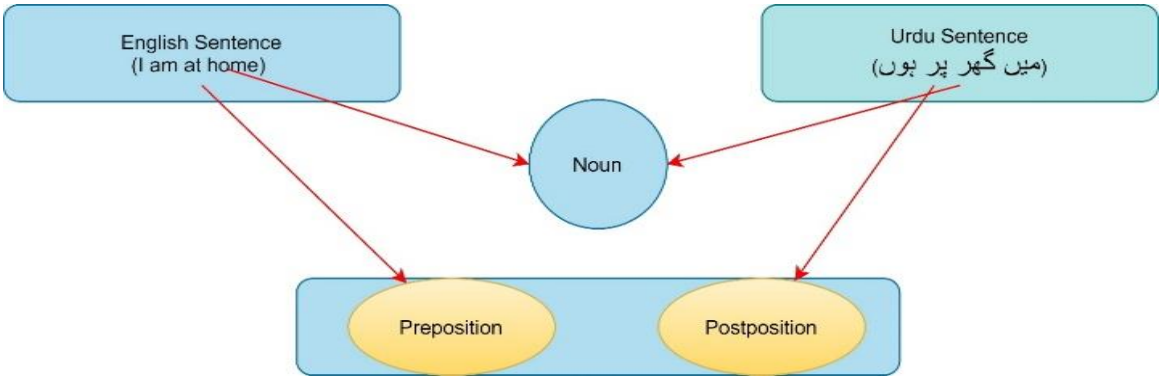

Figure 7. Prepositions and Post-positions in English and Urdu.

Parallel corpus: Parallel corpus is an essential resource for SMT and NMT, and an enormous amount of parallel corpus is necessary for these two approaches. Availability of parallel corpus for low resource languages is also challenging for machine translation.

Sentence Alignment: Sentence alignment is also an essential step in corpus preparation. There are various sentence alignment algorithms and tools available in the literature. The tools available in literature do not support enough for resource-poor languages.

Morphological Variation: Some low resource languages are rich morphological languages in which one word can be inflected in several ways. Translation system should handle all the inflation forms and address forms in training data is challenging.

11. Machine Translation for Low Resource Languages

There are various machine translation systems available in the literature. In this paper, we focused on the machine translation systems available for the low resource languages and their findings. The research work carried out on low resource languages mostly uses direct and rule-based approaches because of the non-availability of massive parallel data to build SMT or the NMT system. In Indian, there are 22 languages given status and official encouragement (8th Schedule of the Constitution). The list of top 15 languages spoken in India is shown in below given Figure 8. It is clear from the figure that Hindi is the most dominant language in India.

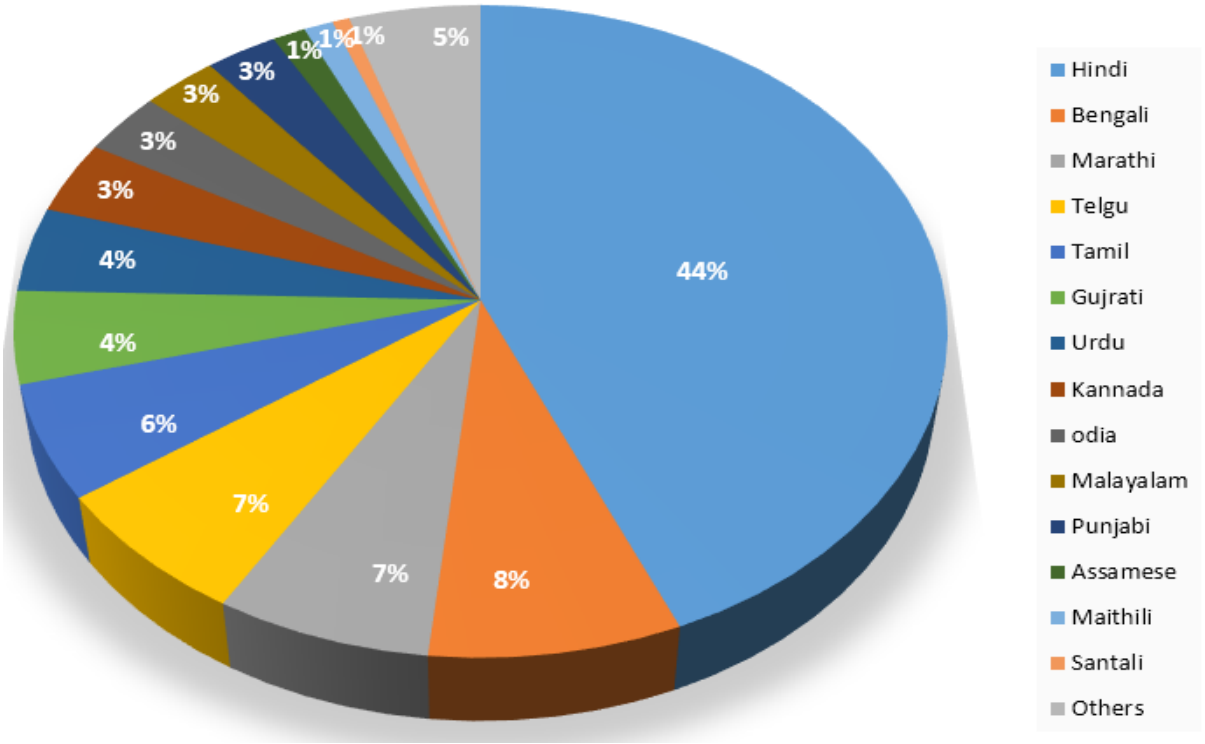

Figure 8. Top 15 languages spoken in India

The below-given Table 3 Various machine translation approaches, the purpose of the machine translation system and findings.

Table 3: Machine Translation System for resource-poor languages

Year	Machine Translation System	Approach Used	Purpose/ Application	Findings
1995	Anusaarka System	(Telugu, Bengali, Kannada, Punjabi, Marathi) to Hindi.	Developed for translating children stories.	Uses Paninian Grammar and matches the similarity between Indian Languages
2001	Anuvaadak Machine Translation English to Hindi [23] [26]	Based on English Hindi Dictionary	General, Agricultural, administrative, and technical purpose	The system uses different inbuilt dictionaries for different domains. The system takes input and identifies prepositions, gender, phrases, and tense of the sentence. After that analysis, the system generates Hindi output.
2001	UNL-Based English Hindi MTS by IIT Bombay. English to Hindi.	Interlingua approach using Use Universal Natural Language as Interlingua for translation	General	Source language is converted into UNL, and from that UNL representation target text is generated.
2002	English-Hindi Translation [7]	Rule-Based	For weather information	Rule-Based with pre-processing of English and Post-processing
2003	AnglaHindi developed by IIT Kanpur English to Hindi [7]	Rule-Based system with context-free Grammar structural.	General Not tuned to a specific domain.	English to Hindi Version of ANGLABHARTI and is web-based. Attempt to integrate rule-based and example-based approach It can be further be enhanced by using the domain-specific parallel corpus.
2004	English to Bangla	Transfer Based Approach	General	The proposed architecture has five stages. It does not perform semantic analysis.
2004	OMTrans English to Oriya [28]	Based on the semantics of source and target language and Grammar	General	Developed used Object-oriented approach and handled word sense disambiguation
2004	Shiva MTS by IISC Bangalore and IIIT Hyderabad English to Hindi	Example-Based Approach	General	Requires a vast amount of parallel corpus and works at phrase and sentence level
2004	Shakti machine Translation IISC Bangalore and IIIT Hyderabad English to Hindi, Marathi, Telugu [26] [28]	Use a Combination of Rule-Based and Statistical Machine Translation Approach	General	The system consists of nine modules for analysing the source text.
2004	English–Telugu Machine [27]	Transfer based	General-purpose applications	Rule-based and can handle complex sentences as well.
2004	Hinglish (Hindi to English) MT System. [26]	Example based approach based on Anubharti and Anglabharti		Implemented on the bases of AnglaBharti II
2005	Anubaad Hybrid MTS English to Bengali by CDAC Kolkata [28]	Hybrid Approach	Translation of News Headlines	The system works at a sentence level.

2006	English to Hindi, Kannada, Tamil Example based MTS	Example based machine translation system with resources as Bilingual dictionary, Phrase dictionary and Word dictionary	General	The system consists of three dictionaries with parallel corpus od words, phrases and sentences the dictionaries are word dictionary, Phrase dictionary, and sentence dictionary The example base consists of 75000 sentences that were manually translated into three languages.
2006	English to Bangla [29]	Example-Based Approach	General	The proposed method uses a shallow analysis to identify input phrases and get target phrases using EBMT.
2007	Punjabi to Hindi MTS	Direct word to word translation approach		The system uses pre-processing modules and performs morphological analysis of source language. This system also performs transliteration. Claimed to achieve accuracy of 92.8% accuracy.
2009	English-Kannada machine aided TS by University of Hyderabad [26]	Transfer Based Approach	Government	The system is based on the transfer-based approach and uses Universal Clause Structure Grammar.
2009	Samparak MT System	Analyse and transfer based paradigm. Rule-based and SMT combination	General	It was developed jointly by 11 institutes and was funded by "TDIL program of Ministry of Electronics and Information Technology (MeitY), Govt. of India". It contains 13 modules to form a hybrid system.
2009	Bengali to Hindi MT [26]	Hybrid approach SMT and Rule-based Approach		
2010	English to Bangla	Phrase-based model of SMT	General	This architecture consists of Prepositional Module, Postpositional module to handle prepositions and post-positions. This architecture also contains a transliteration module, and it overcomes the problem of transfer based approach, but still, there is a problem of ambiguity
2010	English to Urdu [32]	SMT	General	This paper explains the issues regarding the development of parallel corpus and sentence alignment. The tool was developed for automatic sentence alignment.
2010	English to Urdu KBMT [3]	Knowledge-based using Data Mining Techniques	General	Only Proposed Modification and proposed output are mentioned. It is not implemented in real.
2010	Web-Based HPMTS [33]	Direct approach	News, webpages, and also provides e-mail facility	The system architecture has 11 modules Training Module, Unicode Conversion, Normalization, finding and replacing collocations and name entities, translation, ambiguity resolution, transliteration, post-processing, improvement module and test module.
2011	Translation Rules and ANN-based model for English to Urdu	Hybrid (ANN & Rule-based)	General	This system uses a combination of Feed Forward Back Propagation Neural Network with a rule-based approach.

	MT [1]			The knowledge base of Neural Networks is used to store rules and contains a knowledge base of the bilingual dictionary. This system is implemented in Java and Matlab. For training "Levenberg Marquardt backpropagation algorithm" was used. The BLEU score achieved was 0.69.
2012	Malayalam to English MT	Transfer based approach	General Purpose	This system contains a morphological parser for context disambiguation and pre-processor for splitting the compound words. It Contains Bilingual dictionary. This system can be extended to other language pairs
2013	Urdu to English MT using BLEU [4] [7]	Corpus-based approach	General	Mentions paradigms of Machine Translation and their advantages and disadvantages. Compare RBMT, EBMT, Google and Bing using BLEU. Problem is only seven sentences were taken for comparison. EMBT gives good BLEU score of 0.84 i:e 84 %
2013	English to Urdu [32]	Statistical MT approach	General	Uses Moses decoder and IRSTLM for language modelling
2013	English to Malayalam	Hybrid approach combination of SMT with Translation Memory	General	Extends the baseline SMT approach and uses translation memory to remove redundant translations
2014	English to Marathi [36]	Hybrid approach	Mainly focussed on medical reports, tourism agricultural and some web pages	The architecture contains six layers in which each layer perform its task Also compared the output of SMT. RBMT and Hybrid approach proposed and claimed that the hybrid approach is best among the three.
2014	English to Hindi [37]	EBMT Approach	Cannot be used for general purpose as it is trained on 677 sentences	This approach depends on the database, and only 677 sentences used for training
2016	Urdu to Punjabi [25]	SMT	General	The incremental machine learning process was used. Uses Naïve Bayes model for the classification of the input text. Viterbi algorithm was used in the decoding process, and Hidden Markov Model was used as learning Model
2019	Marie: English to Assamese [38]	SMT	Tourism domain as corpus used was from Tourism domain	The results of this system are not so good, as its BLEU score is 0.21. The increase in length decreases the quality of translation

12. Comparison of Machine Translation Approaches

We compared the machine translation approaches based on some basic parameters described below, given table 4.

Table 4: Comparison of MT approaches on basic Parameters.

MT Approach	Cost-Effective	Knowledge from Corpus	Linguistic Background	Mathematical Foundation	Easily Extendable	Reduces Human cost
RBMT	×	×	✓	×	×	✓
SMT	✓	✓	×	✓	✓	✓
EBMT	×	✓	✓	×	×	✓
NMT	✓	✓	×	✓	✓	✓

13. Discussion

In this research paper, several machine translation approaches were mentioned. Many translation systems developed so far mainly used the classical approaches for resource-poor languages. However, little focus was on promising approaches, which are SMT and NMT. Researchers do not apply SMT and NMT due to the unavailability of the enormous amount of the parallel corpus for resource-poor languages. The solution to the problem is to develop dataset using crowdsourcing. We can develop google forms and ask a respondent knowing the language to enter three sentences for each domain like health care, day to day life, tourism, business, etc. The google form will be circulated to the engineering colleges and universities of a particular state. The language is official where students have an e-mail id and awareness about google forms. The second method is that we can create groups on social networking sites for data collection. The other technique is to obtain data from news API's and web scraping, which we have done and have collected 2000 sentences and translated them into the Urdu language using existing translation tools and language experts. We will soon place these parallel sentences in public domain.

14. Conclusion

In this paper, we reviewed various machine translation systems. We found a new promising approach of machine translation like Neural Machine Translation is not applied due to the unavailability of the enormous parallel corpus for resource-poor languages. We also found that some NMT techniques like sequence to sequence model at character level cannot be applied for Spanish and German languages due to complex script writing of some resource-poor language. The long sentence also creates problems in word rearrangement. This paper mentioned machine translation approaches and their evolution, need, challenges of machine translation, and problems faced by researchers in resource-poor languages.MT strives to fill the language barrier gap, and a lot of work has been carried out on European languages. However, Asian languages received less attention. Hence, to fill the language gap, we should try to use promising machine translation approaches to these Asian languages and create some languages.

References

1. Mishra, V., & Mishra, R.B. (2010). ANN and Rule based model for English to Sanskrit Machine Translation. *INFOCOMP Journal of Computer Science, 9*(1), 80-89.
2. Raza, A.A., Habib, A., Ashraf, J., & Javed, M. (2017). A review on Urdu language parsing. *Int. J. Adv. Comput. Sci. Appl, 8*(4), 93-97.
3. Tahir, G.R., Asghar, S., & Masood, N. (2010). Knowledge based machine translation. In 2010 *International Conference on Information and Emerging Technologies*, 1-5.
4. Malik, A.A., & Habib, A. (2013). Urdu to English machine translation using bilingual evaluation understudy. *International Journal of Computer Applications, 82*(7).
5. Okpor, M.D. (2014). Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI), 11*(5), 159.
6. Antony, P.J. (2013). Machine translation approaches and survey for Indian languages. In *International Journal of Computational Linguistics & Chinese Language Processing*, 18(1).
7. Godase, A., & Govilkar, S. (2015). Machine translation development for Indian languages and its approaches. *Int. J. Natl. Lang. Comput. (IJNLC), 4*(2), 55-74.
8. Burch, C. (2001). A survey of machine learning. A survey for the Pennsylvania Governor's School for the Sciences.
9. Tripathi, S., & Sarkhel, J.K. (2010). Approaches to machine translation.
10. Saini, S., & Sahula, V. (2015). A survey of machine translation techniques and systems for Indian languages. In 2015 IEEE *International Conference on Computational Intelligence & Communication Technology*, 676-681.
11. Unlee, P., & Seresangtakul, P. (2016). Thai to Isarn dialect machine translation using rule-based and example-based. In 2016 *13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1-5.

12. Zafar, M., & Masood, A. (2009). Interactive English to Urdu machine translation using example-based approach. *International Journal on Computer Science and Engineering, 1*(3), 275-282.
13. Babhulgaonkar, A.R., & Bharad, S.V. (2017). Statistical machine translation. In 2017 1st *International Conference on Intelligent Systems and Information Management (ICISIM),* 62-67.
14. Imam, A.H., Arman, M.R.M., Chowdhury, S.H., & Mahmood, K. (2011). Impact of corpus size and quality on English-Bangla statistical machine translation system. In 14th *International Conference on Computer and Information Technology (ICCIT 2011),* 566-571.
15. Datta, D., Mishra, S., & Rajest, S.S. (2020). Quantification of tolerance limits of engineering system using uncertainty modeling for sustainable energy. *International Journal of Intelligent Networks, 1,* 1-8. https://doi.org/10.1016/j.ijin.2020.05.006
16. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473.*
17. Gheini, M., & May, J. (2019). A universal parent model for low-resource neural machine translation transfer. *arXiv preprint arXiv:1909.06516.*
18. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144.*
19. Su, J., Zeng, J., Xiong, D., Liu, Y., Wang, M., & Xie, J. (2018). A hierarchy-to-sequence attentional neural machine translation model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26*(3), 623-632.
20. INTERNET WORLD USERS BY LANGUAGE Top 10 Languages. https://www.internetworldstats.com/stats.htm, 20/09/2019.
21. Muhammad, U., Bilal, K., Khan, A., & Khan, M.N. (2008). Aghaz: An expert system based approach for the translation of English to Urdu. *International Journal of Social Sciences, 3*(1), 70-74.
22. Turner, A.M., Brownstein, M.K., Cole, K., Karasz, H., & Kirchhoff, K. (2015). Modeling workflow to design machine translation applications for public health practice. *Journal of biomedical informatics, 53,* 136-146.
23. Jawaid, B., & Zeman, D. (2011). Word-order issues in English-to-Urdu statistical machine translation. *The Prague Bulletin of Mathematical Linguistics, 95*(1), 87-106.
24. Andrabi, S.A.B., & Wahid, A. (2019). Sentence Alignment for English Urdu Language Pair. *International Journal of Recent Technology and Engineering, 08,* 1867-1870.
25. Singh, U., Goyal, V., & Lehal, G.S. (2016). Urdu to Punjabi machine translation: An incremental training approach. *International Journal of Advanced Computer Science and Applications, 7*(4), 227-237.
26. Dwivedi, S. K., & Sukhadeve, P.P. (2010). Machine translation system in Indian perspectives. *Journal of computer science, 6*(10), 1111-1116.
27. Dasgupta, S., Wasif, A., & Azam, S. (2004). An optimal way of machine translation from English to Bengali. In *Proc. 7th International Conference on Computer and Information (ICCIT),* 648-653.
28. Garje, G.V., & Kharate, G.K. (2013). Survey of machine translation systems in India. *International Journal on Natural Language Computing, 2,* 47-65.
29. Naskar, S.K., & Bandyopadhyay, S. (2006). A phrasal EBMT system for translating English to Bengali. In *Satellite Workshop,* 69.
30. Josan, G.S., & Lehal, G.S. (2008). A Punjabi to Hindi machine translation system. In Coling 2008: *Companion volume: Demonstrations,* 157-160.
31. Islam, M.Z., Tiedemann, J., & Eisele, A. (2010). English to Bangla phrase-based machine translation. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation.*
32. Ali, A., Siddiq, S., & Malik, M.K. (2010). Development of parallel corpus and English to Urdu statistical machine translation. *Resource, 10*(5), 31-33.
33. Goyal, V., & Lehal, G. S. (2010). Web based Hindi to Punjabi machine translation system. *Journal of emerging technologies in web intelligence, 2*(2), 148-151.
34. Nair, L.R., & Peter, S.D. (2011). Development of a rule based learning system for splitting compound words in Malayalam language. In 2011 *IEEE Recent Advances in Intelligent Computational Systems,* 751-755.
35. Nithya, B., & Joseph, S. (2013). A hybrid approach to English to Malayalam machine translation. International Journal of Computer Applications, 81(8).
36. Salunkhe, P., Kadam, A.D., Joshi, S., Patil, S., Thakore, D., & Jadhav, S. (2016). Hybrid machine translation for English to Marathi: A research evaluation in Machine Translation: (Hybrid translator). In 2016 *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT),* 924-931.
37. Sinhal, R.A., & Gupta, K.O. (2014). A pure EBMT approach for English to Hindi sentence translation system. *International Journal of Modern Education and Computer Science, 6*(7), 1-8.

38. Liu, J., Zhang, X., Tian, X., Wang, J., & Sangaiah, A.K. (2020). A novel domain adaption approach for neural machine translation. *International Journal of Computational Science and Engineering, 22*(4), 445-453.