JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN COMPUTER ENGINEERING

# SURVEY ON VARIOUS APPROACHES TOWARDS MACHINE TRANSLATION BETWEEN MALAYALAM AND ENGLISH

[1]*ANISREE P. G.*, [1]*RADHIKA K. T.*

[1, 2] **Department of Computer Science and Engineering,**
**M. E. A. Engineering College, Vengoor P. O., Perinthalmanna,**
**Malapuram- 679325 , Kerala, India.**

***anisreepg7@gmail.com***

***ABSTRACT :*** *Machine Translation is an earliest application of Natural Language Processing in which speech or text from one natural language can be translated to another language. The language from which it is translated is called the source language and the language to which it is translated is called the target language. The various approaches towards machine translation are the rule base approach, statistical based approach and hybrid approach. In the first approach the rule of the source language and the target language are taken into account for developing a machine translator, but in statistical approach a set of training data is used for the purpose of translation. Rule based approach is the most risky and time consuming approach because it takes years of efforts to develop a translator by using the complete set of rules. The second approach is the commonly used translator mechanism but it also requires a huge amount of data for developing an efficient system. On combining the qualities of these two approaches a novel approach is developed called the hybrid approach. Studies have shown that this is the most efficient mechanism for developing a translator.*

***KEY WORDS*** *: NLP, MT, SMT, RBMT, HBMT.*

## 1. INTRODUCTION

Machine Translation, perhaps the earliest NLP application, is the translation of text from one natural language to another, using computers. It is one of the interesting and the hardest problem in the field of NLP. India is a multilingual country, i.e., many of the states have their own native language and only 5 percent of the population knows to speak in English. So, it must require a translator which is capable of translating from their native language to English and vice versa for efficient communication and knowledge sharing. The input to the translator is known as source language and the output is called the target language, i.e., for a Machine Translator there is a translation from source language to target language. The research scenario in India is relatively young and machine translation gained momentum in India only from 1980 onwards and various translators are developed for Indian language to English, English to Indian languages and Indian language to Indian language.

The two challenges in Machine Translation are adequacy and fluency. The former is to develop a system that adequately represents the ideas expressed in the source language into the target language. The latter is to represent those ideas grammatically. The common approaches to machine translation are the rule based approach, corpus based approach and hybrid approach. In the rule based approach, a large number of rules are necessary to capture the phenomena of natural language. These rules transfer the grammatical structure of the source language into target language. As the number of rules increases, the system becomes very complicated. Formulation of a large number of rules is a tedious process and require years of effort and linguistic analysis. In the second approach, large parallel and monolingual corpora are used as source of knowledge. This approach can be further divided into statistical approaches and example based approach.

Statistical machine translation is superior to rule based and example based systems in that they do not require human interpenetration and can build a translation system in an unsupervised manner directly from the training data. Rule based systems are language dependent and require careful analysis of source and target languages. With the rapid proliferation of internet and increasing availability of data, SMT is currently the most popular and prevalent paradigm. For an SMT system, a parallel corpus consisting of source and target language sentences and a monolingual corpus consisting of target language sentences are required. The SMT system is trained on these large quantities of parallel data and monolingual data. The statistical model learns the translation parameters from the corpus and performs the translation.

Hybrid based approach is a combination of both the rule based and statistical based approach. It uses a small set of parallel data along with some of the simple rules. Since it uses the rules of both the source and the target language this particular approach could yield high result as compared to any other machine translation mechanisms.

## 2. LITERATURE SURVEY

### 2.1 Example Based Machine Translation System

This is a Machine Translation system [1] for translation from Malayalam to English language. The translation system is based on Example Based Machine Translation approach. Example Based machine translation is based on the idea of reusing the already translated examples. It involves three major steps - Example acquisition, Matching and Recombination as shown in the following figure.

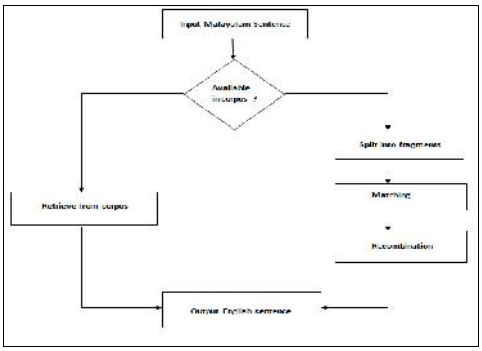


Figure 1: EBMT System Architecture

**Example Acquisition**
Example acquisition is the process of acquiring examples of already translated sentences and to form a parallel corpus for the translation system. Corpus is the collection of the examples from various resources. In this work, corpus not only contains the examples but also idioms and collocations, multiword terminology and phrases are included. The work mostly uses idioms, multiwords and phrases in Malayalam language and its corresponding translation.

**Matching**
Matching phase is the one of the major steps in Example Based Machine Translation. Corpus is searched for finding out the best matching for the input source sentence. Also it deals with how these stored examples are used for the translation. Sometimes it is very difficult for the system to translate a full sentence in itself. Then the input sentence is split into smaller fragments. In this case, first look at the example database (corpus) and find out the longest possible fragment available in the corpus and select the corresponding translated fragment. Then, consider the remaining part of the input sentence for which the next matching fragment has to be found from the corpus. This process will continue till the end of input sentence. If the system does not have extensive corpus, matching process may not be successful.

**Recombination**
This is the last module. Here the fragmented sentences are recombined to form the output sentence. Hence the recombination enhances the readability of the target sentence. Combining these translated chunks into a well formed structure in the target language is the most difficult step in EBMT. But it has received always less attention than all the other steps in translation.
They tested the system with different kinds of sentences in Malayalam language. This Example Based Malayalam to English translation system generates correct meaningful English sentence as output in most of the cases. The system works well for the all simple sentences in their 9 tense forms, their negatives and question form. Evaluation of this translation system was done by them manually. Quality of the translation is measured by how perfect the translated sentence in English. According to their study about 75 percent of the test set yields good quality translation. The translation system completely relies on the corpus that contains examples of already translated words, phrases and sentence.

### 2.2 Transfer Based Machine Translation System

This paper [2] describes a transfer based scheme for translating Malayalam, a Dravidian language, to English. The system comprises of a preprocessor for splitting the compound words, a morphological parser for context disambiguation and chunking, a syntactic structure transfer module and a bilingual dictionary. All the modules are morpheme based to reduce dictionary size. The system uses two sets of rules: rules for Malayalam morphology and rules for syntactic structure transfer from Malayalam to English.

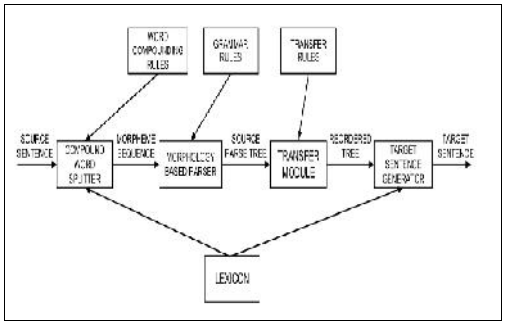


Figure 2: Transfer based machine translation system

**Compound Word Splitter Module**
Morphological variations for words occur in Malayalam due to inflections, derivations and word compounding. The compound words are to be separated before any further processing. Splitting has been done at morpheme level to reduce dictionary space. The sequence of morphemes is given to the

parser for chunking and word sense disambiguation. Due to the ambiguity in the splitting rules the system generates multiple splits for the same input sentence and the split with least number of constituents is fed to parser.

**Parser Module**

Parser takes output from the splitter and does the following tasks. It groups the input sequence of morphemes into chunks and performs word sense disambiguation based on morpheme tags. The chunking process finds the basic units for tree reordering. The parser uses a depth first approach with backtracking. The output of the parser is a parse tree for the next module. The parser uses the syntax rules for the morpheme sequences in Malayalam sentences in the regular expression form.

**Syntactic Structure Transfer Module**

The transfer module transfers the source language structure representation to a target language representation. This module needs the subtree rearrangement rules by which the source language sentence syntax tree can be transformed into target language sentence syntax tree. The system performs most of the commonly needed reordering for Malayalam to English translation.

**Target Sentence Generator Module**

The generation module generates target language text using target language structure. This uses inter chunk dependency rules and intra chunk dependency rules. It involves lexical transfer of verbs, transfer of auxiliary verb for tense, aspect and mood and transfer of gender, number and person information.

**Cross Lingual Dictionary**

The dictionary includes most of the commonly occurring verbs, nouns, pronouns, adjectives, inflectional and derivational suffixes, clause suffixes etc. Each entry in the file has three fields: the root word (morpheme), the morpheme tag and its translation. The verbs in past tense have their root words stored along with them. Since the system works with morphemes, the space required for the dictionary is less. The system works for sentences which contain upto two adverbial or adjectival clauses which is commonly found in Malayalam texts. They said that the system can be modified to handle other sentences by adding appropriate grammar rules and transfer rules to the rule database. As the parser is a general parser, it can handle sentences of any depth. In around 20% of sentences to the system returned the exact English version of the input sentences.

**2.3 Machine Translation Using Hybrid Approach**

This is a machine translation [3] from English to Malayalam using a hybrid approach. A hybrid approach is always a combination of statistical machine translation and a rule based machine translation. Here it is named as hybrid in the sense that it extent the statistical approach with a translation memory, where the translation memory is used as a cache which store the recent translation and hence avoid redundant translations. So this system consist of two parts, they are statistical machine translator and translation memory as shown in figure.

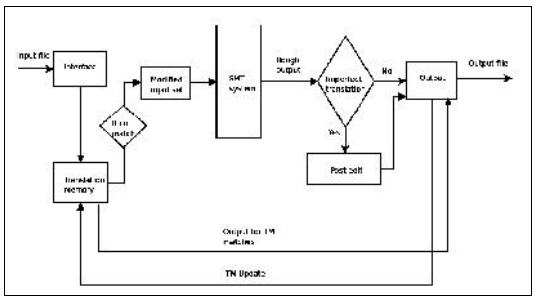


Figure 3: Hybrid Approach to Machine Translation

**Statistical machine translator**

This module uses the pure statistical approach of the machine translation.

SMT can be split into two phases:

- Training phase
- Translation phase

In training phase the system trains how to translate and in translation phase the system translate the input sentence to the output sentence. The training phase can be again divided into three major steps. They are:

- Collecting the quantitative data for training, both the monolingual and bilingual corpus
- Build the language model for the target language from the monolingual corpus
- Build the translation model from the bilingual corpus

The translation phase uses a heuristic search for identify the good translation for input source sentence.

**Translation memory**

The major idea of using a Translation Memory in this project is that, most of the translations are repeating in nature. If it is possible to find the existing translation then the redundant translation can be avoid. Here, translation memory is act as a cache, in which the previous translations are stored. TM analyzes the input sentence and checks if it is already available in its database. If it finds a match it will use the corresponding translated output. An unseen source sentence is handled by the machine translator itself. Thus TM eliminates the effort for running the translator for a previously translated sentence.

A TM contains four main components.

- A mechanism to store sentences and their translations

- A search mechanism to find input sentence matches from TM
- Provision for post editing the translator output
- Provision for updating the TM

They evaluate the system both in manual and with BLEU score. A total of 70 input sentences is given to the system and a BLEU score of 69.33% is obtained.

### 2.4 Statistical Based Machine Translation

This is an English to Malayalam machine translation paper [4] which uses a statistical based approach. In the training process the translations of a Malayalam word is determined by finding the translation probability of an English word for a given
Malayalam word. They collect the corpus data from online Malayalam newspapers and magazines. Since it requires a word by word translation of the bilingual corpus, it is very much risk to develop the same. Also, it is very much risk to find the translation from the corpus. The reason behind this occurrence is solely the peculiarity of Malayalam language. A linguist when asked to translate sentences into Malayalam, have a wide range of options to apply.

**POS tagging the bilingual corpus**

The word by word translation is performed with alignment models. Alignment is a process of mapping English word with Malayalam words. So the total number of alignments is depends upon the total number of words. Since the Malayalam word are suffix separated the number of words also get increased, this further increase the number of translated output. This could be overcome with POS tagging. So, the same category words only align together. This will reduce the number of translations.

**Suffix separation from Malayalam corpus**

Malayalam language is enriched with enormous suffixes. The suffix separator is employed to extract roots from its suffixes. Suffix separation rules are formed by applying sandhi rules in Malayalam in the reverse direction.

**Stop word elimination from the bilingual corpus**

The Malayalam corpus after suffix separation will contain many suffixes extracted from root words that have no meaningful word translation in English. Since these words are useless in the translation process, they may not be included in the corpus. The deletion of these stop words will bring down the complexity of the training process as well as improve the quality of the results expected from it. Similarly, stops words in English language are also identified and are eliminated from the corpus before subjecting it to training.

**DECODING PHASE**

For the decoding, they use Bayes Rule. The outcome of the decoder is further modified for better output. The steps are given below:

**Tagging the English sentence**

In the decoder different syntactic tags are used to denote the syntactic category of English words.

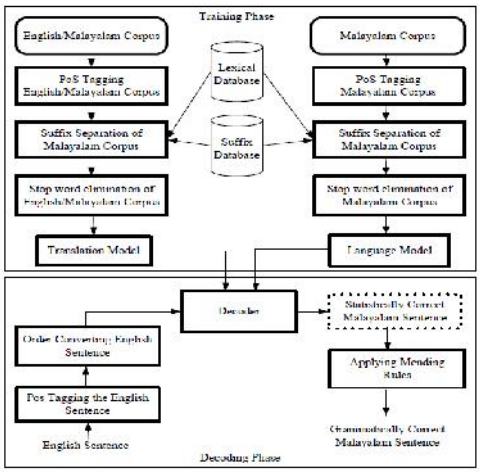


Figure 4: Statistical based Machine Translation

**Order conversion**

Since English and Malayalam belong to two different language families, they totally differ in their subject verb order. Order conversion rules are framed to reorder English according to the sentence structure and the word group order of Malayalam.

**Generating Statistically Correct Malayalam**

The order converted English sentence is split into phrases and a phrase translation table with different options of Malayalam translations is developed. Various hypotheses are created by choosing translation options and the best translation is determined by extending the hypotheses and picking the one with maximum score.

**Generating Grammatically Correct Malayalam**

SCM fails to convey the complete meaning depicted in a sentence. This undesirable result has been set right by applying various mending rules which helps in converting SCM into GCM.
For an unseen sentence the baseline method with suffix separation gives a BLEU score of 0.38. Even though the translation produced depicts correct meaning of the English sentence, they said that the expected score is not met. This is due to the large number of word substitutions rather than insertions and deletions occurring in the translated sentence when compared to the reference text.

### 2.5 Rule Based Machine Translation

This is a translation paper [5] from English to Malayalam using the rule based approach towards machine translation. The core process is done with bilingual dictionary of English-Malayalam pair and

rules for converting source language structure to the target language structure. There are mainly two types of rules are used by them, one is transfer link rules and the other one is morphological rules. Where the transfer link rules are used for obtaining target structure and morphology rules are used for assigning morphological features.

This rule based machine translation from English to Malayalam works with the help of five files. They are ROMANTOUNICODE file, UNICODETOROMAN file, word dictionary file, morphdictionary file and transfer link rule file. The word dictionary includes all the verb, noun, determines, adjectives etc. It contains the English words, root word, POS tag information, the corresponding Malayalam word and the exact tense for this Malayalam word. Morphdictionary file has the information about the source features, POS tags and target features. Transfer link rule file gives information about how the target language sentence looks likes. It is very important for an English to Malayalam translation because for English it follows Subject, Verb, Object format but for Malayalam it is Subject, Object, Verb.

ALGORITHM

- Step 1: Extract and store the tokens in the sentence
- Step 2: The source sentence is given to parser. The POS tag information, source tree structure and source dependency information obtained from the parser are taken
- Step 3: By considering POS tag information and source dependency information, source morphological feature is assigned
- Step 4: The target tree structure is generated by considering the Transfer Link Rule File
- Step 5: The word order is generated by considering the source structure and target structure
- Step 6: By considering the source morphological feature and POS tag information, target morphological feature is extracted from Morphdictionary File
- Step 7: The word by word translation is done based on target morphological feature, by using mapping files and word dictionary file
- Step 8: Finally, by using word order, translated words and target tree structure, translated output sentence in Malayalam is generated

In short, in this rule based technique, firstly, the words from the source sentence are taken separately. The POS tag information and dependency information of these words is obtained with the help of a parser. Using this information, source morph features are assigned to each word. Then the corresponding target structure is generated with help of transfer link rule file. Finally, with the help of word dictionary and Morphological dictionary, the target sentence is generated. The major drawbacks of the system are:

- It can translate the source sentence with word limit of six
- If an English word contains more Malayalam meaning it translate more sentences

### 2.6 Syntactic Based Machine Translation

This SBMT [6] system is specifically designed for translating text in English to Malayalam. For the translation purpose this system uses a bilingual English-Malayalam dictionary and a morphology generator. This is basically a transfer based machine translation paper. The basic idea is rearranging the nodes in the source language tree to the target language tree. General rules are identified for certain sentences and these rules are used for translating new sentences. This system mainly consists of four modules. They are:

- Syntax tree generator: Preprocessing of the input text is carried out in this module
- Word Reordering: Converts SVO form to SOV
- Pattern Recognition: Identify the sentence pattern of the text
- Translation: Translate the reordered English text to Malayalam text

**Syntax Tree Generator**

Here Stanford parser is used for identify the POS tag and dependency information of the source sentence. POS tagging is the process of assigning a part-of-speech like noun, verb, pronoun, preposition, adverb, adjective or other lexical class marker to each word in a sentence. Dependency information represents dependencies between individual words.

**Word Reordering**

Different languages have different syntactic structure. For example English follows a Subject, Verb, Object format whereas Malayalam follows a Subject, Object, Verb format. And also the main verb is always in the last for Malayalam. This shows the importance of word reordering. For most of the cases word reordering is performed with statistical processes, so it does not requires any syntactic information. Here, they use a new approach for syntactic transfer. In order to reorder, different phrases are extracted from the syntax tree. Phrases are considered as the second level of classification as they tend to be larger than individual words, but are smaller than sentences. Different phrases constitute a sentence.

**Pattern Recognition**

In this module various sentence patterns are identified based on the dependency information generated by the parser. General rules or tag sequences are identified based on the POS Tag information of the reordered sentence. According to each tag sequence certain case information is added to the POS tag.

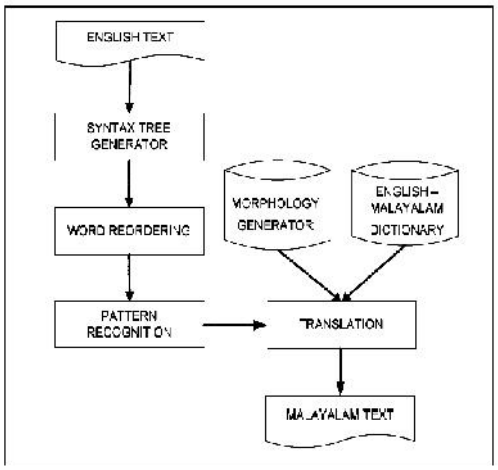


Figure 5: Syntactic based machine translation

**Translation**

Reordered sentence are translated using a bilingual English-Malayalam dictionary. The translation also uses the morphology generator for noun, pronoun and verb. Morphology generator analyses the internal structure of the translated text. Malayalam is an agglutinative language in which a word is formed by adding suffixes to the root. Nouns are linguistic categories, which takes cases and PNG (Person, Number and Gender) information. Nouns change their forms according to Case (Vibhakthi), verbs change their forms according to the TAM (Tense, Aspect, Mood). Person can be first person, second person or third person. Number is a part in an utterance, which indicates whether on object we talk about, is one, or more than one. Gender in language is the same as the universally known divisions of masculine, feminine and Neuter.

Performance of the SBMT system was measured by them using Word Error Rate as well as the F-measure. For an unseen sentences the WER is 0.333 and the corresponding F- measure is 0.66. The major drawback is that the size and quality of dictionary limits the scope and coverage of the system.

## 3 OBSERVATION AND RESULT ANALYSIS

The following table shows the survey result of various papers. From the table, it is very much clear that if we are using a hybrid approach towards machine translation, the result is very much high.

| PAPER | APPROACH | RESULT | REMARK |
|---|---|---|---|
| A Hybrid Approach To English to Malayalam Machine Translation | Hybrid Based | 69.33% BLEU score | The system result a good score and human post editing can also be used here |
| English to Malayalam Translation: A Statistical Approach | Statistical Based | 38% BLEU score | The expected score is not met due to large number of word substitutions when compared to the reference text |
| Syntactic Based Machine Translation from English to Malayalam | Transfer Based | 33% WER 66% F-measure | The size and quality of dictionary limits the scope and coverage of the system |
| Rule Based Machine Translation from English to Malayalam | Rule Based | Average | It can translate the source sentence with word limit of six and for the English words that contains more Malayalam meaning it translate more sentences |
| Design and Development of a Malayalam to English Translator: A Transfer Based Approach | Transfer Based | 20% | Rich set of Malayalam rules are required for this system. There are some shortcoming due to the incomplete use of rules |
| Malayalam To English Machine Translation: An EBMT System | Example Based | 75% of test data give good result | The system is good for translating Malayalam sentences, but only simple sentences can be translate |

## 4 CONCLUSION

In India, researchers have been pursuing on MT since 1980. Different MT systems has been developed and is using in different parts of India. Out of all the 22 official languages some of the languages is not showing a good result in machine translation and not a tremendous research is focused on these languages. Malayalam is spoken by 38 million people in the south east state Kerala and is one such language. The peculiarity nature of Malayalam is the major reason for this. But we must requires a translator which can translate Malayalam and English. The studies have shown that if we are using a hybrid approach towards the translation between Malayalam and English this could yield high result.

## 5 REFERENCES:

[1] E. S. Anju and K. V. Manoj, “Malayalam to English machine translation : An ebmt system,”

IOSR Journal of Engineering, vol. 4, pp. 18-23, January 2014.
[2] R. Latha, D. Peter, and R. P. Ravindran, "Design and development of a Malayalam to English translator - a transfer based approach," International Journal of Computational Linguistics (IJCL), vol. 3, 2012.
[3] B. Nithya and J. Shibily, "A hybrid approach to English to Malayalam machine translation," International Journal of Computer Applications vol.81, no. 8, November 2013.
[4] S. Mary, K. Sheena, and G. Santhosh, "English to Malayalam translation: A statistical approach," Proceeding of the 1st Amritha ACM-W Celebration on Women in Computing in India, 2010.
[5] R. Remya, S. Remya, R. Remya, and K. P. Soman, "Rule base machine translation from English to Malayalam," International conference on Advances in Computing, Control and Telecommunication Technologies, 2009.
[6] T. Anitha and Sumam, "Syntactic based machine translation from English to Malayalam," International Conference on Data Science and Engineering (ICDSE), 2012.