IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 04, Issue 01 (January. 2014), ||V1|| PP 18-23

Malayalam To English Machine Translation:An EBMT System

Anju E S[1], Manoj Kumar K V[2]

[1]*(Dept. of Computer Science & Engineering ,Govt.Engineering College Thrissur)*
[2]*(Professor,Dept. of Computer Science & Engineering ,Govt.Engineering College Thrissur)*

Abstract: **-** Machine Translation became important in the field of communication around the world with people having their own native language. Machine translation is one of the research areas under computational linguistics. Various methods have been proposed to automate the translation process. The thesis proposes a Machine Translation system for transaction from Malayalam to English language. The translation system is based on Example Based Machine Translation (EBMT) approach. The input to the translation system is Malayalam sentence and the corresponding English sentence is generated as output. Example Based machine translation is based on the idea of reusing the already translated examples. Example based translation involves three major steps - Example acquisition, Matching and Recombination. It is founded that the translation system works well for the simple sentence in Malayalam language.

Keywords: - *Natural Language Processing, Machine Translation, Example Based Machine Translation.*

I. INTRODUCTION

The language is an effective medium for the communication that conveys the ideas and expression of the human mind. There are more than 5000 languages in the world for the communication. To know all these languages is not a solution for problems due to the language barrier in communication. In this multilingual world with the huge amount of information exchanged between various regions and in different languages in digitized format, it has become necessary to find an automated process to convert from one language to another. Natural Language Processing (NLP) is one of the hot area of research that explores how computers can be utilize to understand and manipulate natural language text or speech. In NLP research we are gathering the information about how humans understand and use the languages, then developing tools and techniques to make the computer system to manipulate the natural languages etc. There are many applications of NLP such as machine translation, cross language information retrieval (CLIR), speech recognition, and artificial intelligence and so on. Natural language processing is an ongoing challenging and complex research topic.

Machine Translation is the process of enabling a computer to translate sentences from one language to another. There is lot of research going on this area of machine translation at present. Many works are concentrated on the translation of English to Indian regional languages. Moreover, translating Indian regional languages to English is also important for many applications. In Kerala, most of the population is not so familiar with English. English is an international language and one of the popular spoken languages in the world. The translation of the native language into a commonly used language is essential for many applications like instant message systems and for all communication systems. The thesis proposes a translation system which translates a Malayalam sentence into its corresponding English sentence. There are many different approaches to the machine translation. Example based machine translation(EBMT) is one of the dominant type of machine translation. It is a corpus based approach, based on the idea of translation by analogy. Objective of this thesis is to build an example based machine translation system which translates sentences in Malayalam language into the English language. Input to this translation system is a Malayalam sentence and output its corresponding English sentence. The system works through the three major steps of example based machine translation. The first step is the example acquisition, i.e. creating the parallel Malayalam English corpus, the second step performs matching of input sentence against the corpus and the last step deals with recombination and generation of output sentence.

The next section gives a survey on Machine Translation. The third section describes relevance of example based machine translation (EBMT), the proposed system and various steps in the example based machine translation. Section 4 deals with the implementation aspects. Section 5 describes the results followed by conclusion and future work.

II. LITERATURE SURVEY

This section will give the idea about the significance of the area of machine translation and identify a place where a new contribution could be made. The section deals with published information in the area of machine translation within certain time period. A lot of research has been done in the area of machine translation in different parts of the world. There exist many machine translation systems for the different languages across the world. The history of machine translation began in the 1950s after the second world war. A translation system by a group of researchers from Georgetown University which translates Russian sentences into English language. But in 1966, a famous report (ALIPAC Report) published which stated that machine translation is a work for wasting time and money. This negative report slowed down the further research for a some years. In 1970s, machine translation activity extended from the United States to Canada and to Europe. Many different types of machine translation systems were developed and most of them follow Interlingua approach. The majority of MT research was the rule based machine translation by the end of the 1980s. All the machine translation works are based on linguistic rules for syntactic, semantic and morphological analysis of language. This dominance of the Rule Based Machine Translation approach has been broken by corpus based machine translation methods.

Idea of EBMT system is put forward by Nagao in his famous translation by Analogy paper in 1984[9]. In the beginning of 90s Machine Translation researchers were attracted to statistical and example based machine translation. EBMT became an established field for research in Machine translation by 1993. The great demand for high-quality automatic translation made almost all the researchers move towards the corpus-based machine translation. There are so many works done in this machine translation technique for different language pairs in the world. The section describes the various approaches for machine translation and works related to each approach.

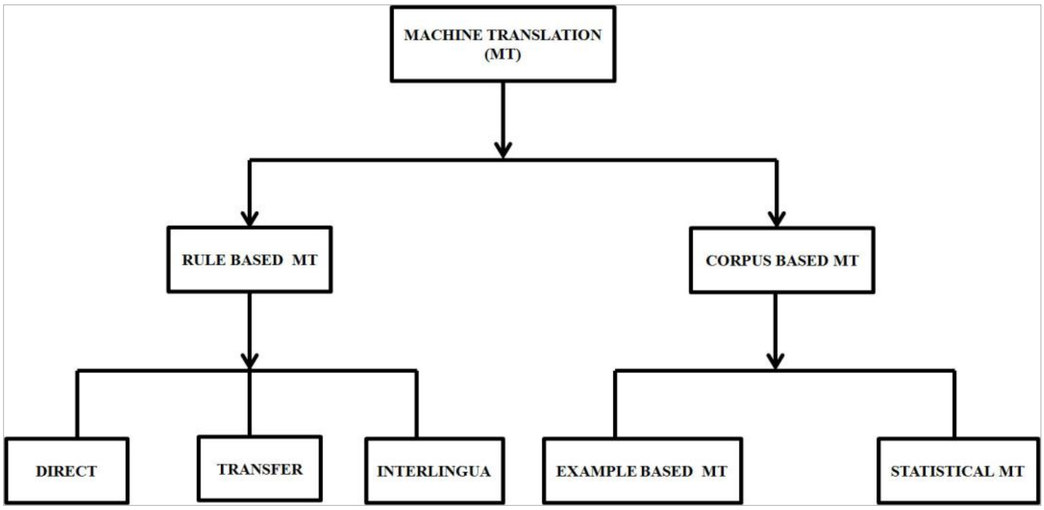

Fig 1: Types of Machine Translation

2.1 Rule-Based Machine Translation-RBMT

Rule-based machine translation (RBMT) uses linguistic rules to analyze the input sentence in source language to generate text in the target language. RBMT mathematically break down the source and target languages using linguistic information. So it is more predictable and grammatically superior than other methods. An RBMT system generates output sentences on the basis of morphological, syntactic, and semantic analysis of both the source and target languages involved in translation task. Usually such systems incorporate linguistic tools such as morphological analyzers, generators, disambiguator and syntactic parser. These tools can create a knowledge rich representation of the input and the output language sentences and produce fluent and grammatical output. But the limited coverage of the vocabulary and syntactic rules with the ambiguity causes deterioration in result quality, and maintaining very large rule based systems become very difficult. According to the nature of the IR used, rule based Machine translation can be divided into three major classes direct, transfer-based and inter-lingua. A direct MT is a direct word by word translation of the source language to target language without passing through an intermediate language construct. Translation highly dependent on both the source and target languages do not focus on the detailed linguistic rules. In transfer based machine translation, the source language is converted into the internal representation (IR) based on the source language. An equivalent representation for the target language is generated from this IR form using grammar rules and bilingual dictionary. In the inter-lingua approach text in the source language is transformed into an intermediary language which is independent of source and target language. This translation system needs only two modules, analysis and synthesis. The system is relevant in multilingual machine translation since it is independent of the language pair.

Rule Based Machine Translation System [4] was developed for English to Malayalam language in 2011. This RBMT system takes English sentence as input and produce corresponding Malayalam sentence using a Stanford Parser. It is purely a rule based machine translation system. The main use of Stanford Parser is the source (English) side processing, Parsing, POS tagging, stemming and Morphological analysis. The English to Malayalam bilingual dictionary and font converters of Malayalam is used for the translation. Another rule based machine translation system for English to Malayalam [7] based on syntactic information about the source language. The Main idea was arrangement of nodes in the parse tree of the source language to the target language. Basically it was a transfer based bilingual translation system. Some general rules were used for particular types of sentence patterns. The translation system comprised of bilingual English-Malayalam dictionary and a morphological generator. Translation quality of the system was improved by Parts Of Speech (POS) tag information of the texts. In 2012 , A transfer based scheme for translating Malayalam[1], to English was developed. It was a rule-based machine translation along with different knowledge components of both languages. The system consists of several modules like a pre-processor, morphological parser, a syntactic structure transfer module and a bilingual dictionary.

2.1 corpus-based machine translation-CBMT

Since 1989, corpus based approach for machine translation has emerged as one of the widely explored area in machine translation. Due to the high level of the accuracy achieved during the translation, this method has dominated over other approaches. Corpus based machine translation is a type of translation which over comes the knowledge acquisition problem of RBMT. It uses a huge bilingual parallel corpus to obtain knowledge for the translation. This translation system transforms source language into form that depends only on the target language. Output text is generated from this intermediate form. CBMT system are classified into Statistical Machine Translation-SMT and Example Based Machine Translation-EBMT

Statistical machine translation is an entirely different approach which never creates such complex linguistics rules like RBMT system. Statistical machine translation is a data-driven machine learning method based on huge bilingual corpora. Here the translation is a probabilistic task that uses statistical rules which learns from parallel corpora for the translation. Google Translate is an example of Statistical Machine Translation (SMT). Google applied different algorithms based on probability and statistics with a wide and extensive corpus to generate translation. The advantage of SMT system is that linguistic knowledge is not required for building them. The difficulty in SMT system is creating massive parallel corpus. Example Based machine translation (EBMT)is based on the concept of recalling or finding analogous examples of the languages, that is translation by ”Analogy”. EBMT system has a set of sentences and already translated equivalent sentences of them in target language. These examples are used to translate the similar type of source language to the target language. Example based machine translation stands somewhere between Rule based machine translation and Statistical machine translation. It has both rule based and data driven features.

In 2010, an SMT system for English to south dravidian language [3] was developed. The paper proposed a statistical machine translation (SMT) system for English to south Indian languages such as Malayalam and Kannada. The translation system works for almost all simple sentences in their twelve tense forms, their negatives and question forms. Conventional statistical machine translation (SMT) sometimes fails to find a good translation due to problems in its statistical models as well as search errors during the decoding process. Pure statistical based machine translation (SMT) sometimes fail to produce quality output for the longer sentences. The rule-based machine translation also face problems like high cost in formulating rules and inconsistencies when the number of rules increase. The performance of an example-based system relies on the parallel corpora and similarity measures between example and input sentences. Integration of these approaches may overcome shortcomings associated with each method .Example-based rescoring of statistical machine translation [8] was developed as an example-based rescoring method that validates SMT translation candidates and judges whether the selected decoder output is good or not. The system uses a validation filter which rejects the defective translations..A linguistics-lite EBMT system was proposed in the thesis Example-Based Machine Translation using the Marker Hypothesis [5] in 2005. The EBMT system is based on the Marker Hypothesis and exploring the application of the Marker Hypothesis in an (English, French) example-based system. This system successfully deduced a set of sub-sentential aligned chunks, words and generalised templates in the system. In 2006, another linguistics-rich EBMT work was reported in the paper Application of Linguistic Rules to Generalized Example Based Machine Translation for Indian Languages [10].This EBMT system also based on the marker hypothesis. In 2009, an approach for English to Malayalam SMT machine translation was proposed. A rule based reordering and morphological processing [6] is basically a SMT system. The work combines rule based reordering and morphological information for the translation. Basic ideas are reordering the English source sentence with respect to Malayalam syntax, and the root suffix separation is used for both English and Malayalam words.In 2007, a work was reported on this concept that A Hybrid Approach to Example based Machine Translation for Indian Languages [11].It was a hybrid approach to example based machine translation making use of statistical machine translation methods and minimal linguistic resources. Very recently in 2013 Harbinger Kaur and Dr. Vijay Laxmi proposed a hybrid approach of RBMT and EBMT. A Web Based English to Punjabi MT System for News Headlines [2] presented English to Punjabi MT system that translates news headlines of an English news paper into Punjabi in a particular domain. The system will help for those who are unaware of English. The system follows different approaches to machine translation such as direct, example-based (EBMT) and rule based.

Lot of research work has been going on the area of machine translation for several decades with different approaches. Some recent studies focussed on a translation between Malayalam and English and works were reported for translation from English to Malayalam. A Rule Based Malayalam to English translation system was reported in 2012 for translating a Malayalam sentence to English. It was the first effort reported in the translation from Malayalam to English. Hence research on Malayalam to English translation is very relevant and necessary. Development of rule-based systems is expensive, time-consuming and required extensive linguistic rules. This work follows a corpus based EBMT approach. Also to our knowledge, there has been no work on Malayalam to English translation system with EBMT approach.

III. PROPOSED SYSTEM

This thesis proposes an example based machine translation system for Malayalam to English. It is the first machine translation work on Malayalam to English translation incorporating EBMT technique. Input to this machine translation system is Malayalam sentence and the output of the system should be its corresponding English sentence. And this system will generate the output English sentence with high quality of translation. The three main steps in Example Based Machine Translation (EBMT) are Example acquisition, Matching and recombination.

3.1 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 contain examples in the sentence level but also in various more interesting levels such as sub-sentential levels including words, idioms and collocations, multi-word terminology, and phrases. The work mostly uses idioms,multi words and phrases in Malayalam language and its corresponding translation.

3.2 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. Matching is concerned with find out the matching fragments in the corpus against the input 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 undergoes splitting and creates a set of smaller fragments. We need to solve the problem of segmenting a given input sentence in case of no translation available in corpus as complete sentence. In this case, we 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, we 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
Let S be the input Malayalam sentence decomposed in to small fragments called examples (e1,e2,e3...)
D(S)=e1,e2,e3,.. Where D indicates the decomposition
Then we look up in the corpus to get the translated fragment for each example
T(e1)=t1
T(e2)=t2
T(e3)=t3 Where T indicates the translation and t1,t2,t3 are the English fragments corresponding to e1,e2,e3 respectively

3.3 Recombination

This is the final step in the example based machine translation. Recombination or sentence synthesis is the process of combining the translated fragments in to target text. Hence the recombination generates the target translated sentence and 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

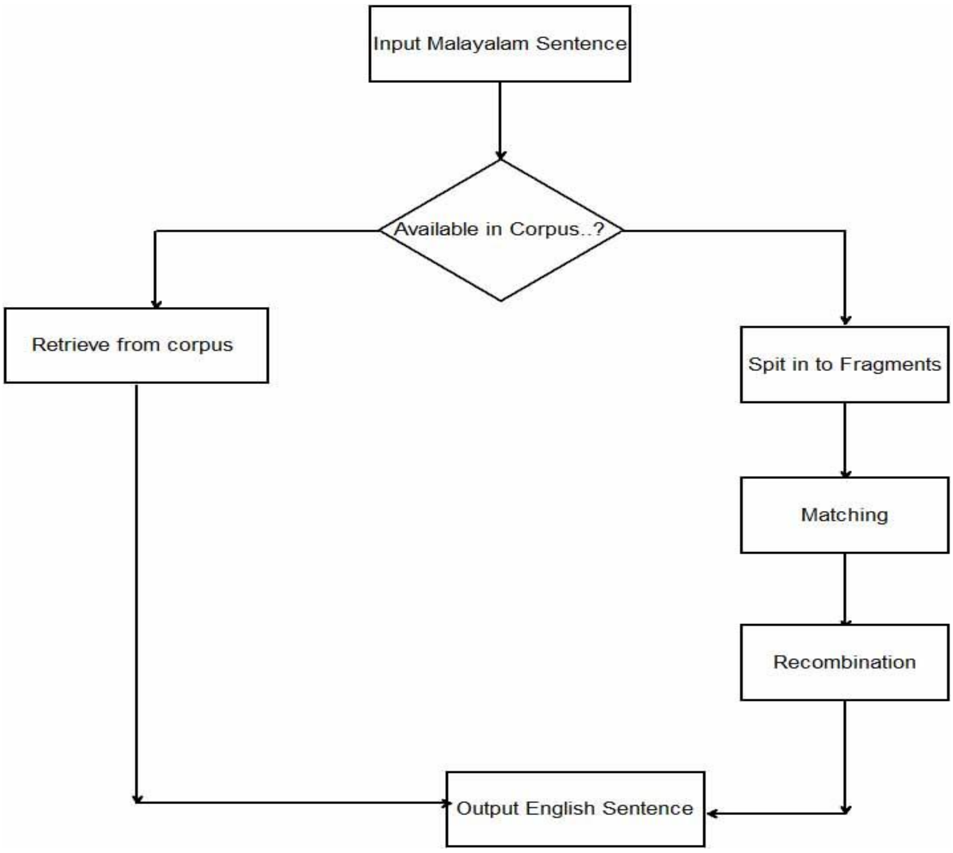

Fig 2:EBMT system

IV. IMPLEMENATION

The proposed work is implemented with MATLAB version 7.10 which links with some java code. Matlab is a language for prototyping ideas. It comes with many libraries especially for machine learning and statistics. The translation system implementation goes through three major steps such as Example Acquisition, Matching and Recombination. In the first stage, it creates a parallel corpus for the system which contains already translated examples. In this translation System the corpus is saved as .MAT file in Matlab. A MAT-file stores data in binary form. Matlab provides application program interface routine for read and write operation on this MAT file. Matlab editor does not support Malayalam fonts, hence processing of Malayalam cannot be possible with Matlab. So the Malayalam data has been converted into Unicode font. The corpus consists of a set of Unicode of Malayalam and its corresponding English. Matching with the corpus is the process of finding the translated fragment from corpus by indexing into MAT file. The system searches in the corpus for each Malayalam fragment and retrieve the corresponding translated English fragment. For string comparison, the system uses Matlab built-in function and also levenshtein distance algorithm. Output of this matching phase is the set of English fragments corresponding to fragmented Malayalam sentence stanford POS tagging is used after matching phase. The Stanford POS tagging written in javacode is linked with Matlab. Recombination makes the output English sentence in theform of SVO based on some rules. So the quality of output of the EBMT depends on the rules written in recombination stage.

V. RESULT

The system was tested with different kinds of sentences in Malayalam language. The prepared test set consists of simple sentences in Malayalam language. The simple sentence which contains only one independent clause and no dependent clauses. 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. A group of sample input sentences with the tabulated outputs are shown (Table 1) below to give a picture of the results obtained. Evaluation of this translation system was done manually. The human experts in translation evaluate the translation quality of this example based machine translation. Quality of the translation is measured by how perfect the translated sentence in English. About 75% of the test set yields good quality translation while the remaining faces some reordering problems. The translation system completely relies on the corpus that contains examples of already translated words, phrases and sentence. System performance can be improved by a large aligned corpus and more rules for reordering.

Table 1: Sample Output

Input Malayalam Sentence	Output English Sentence
ഞങ്ങള് പുസ്തകം വാങ്ങി	we bought book
എനിക്ക് പാല് ഇഷ്ടമല്ല	i do not like milk
അവന് ഉറങ്ങുകയായിരുന്നു	he was sleeping
ഞാന് കളി ജയിക്കും	we will win game
ഞാന് ആംഗലഭാഷ സംസാരിക്കുന്നു	i speak English
അവൻ ആ സിനിമ കണ്ടിരുന്നു	he had seen the film
എന്റെ അച്ഛന് ഡോക്ടറാണ്	my father is a doctor
അവൾ വരാറുണ്ടോ ?	does she came ?
അവര് റോഡിലൂടെ ഓടുന്നു	they are running on the road

VI. CONCLUSION

The work proposed and built a new approach for Malayalam to English translation. The computers are the most suitable machines to remember the things like translated examples. This EBMT system is based on the idea of reusing the already translated sentences or phrases. It successfully translated simple sentence in Malayalam language to English. It can be extended to many applications by slight variations. NLP is one of the hot areas of research nowadays. This research may make an impact in this area especially on machine translation by Malayalam to English translation. There are varieties of application for this translation system. In Kerala, all the population is not so familiar with English. So such kinds of systems will offer a great contribution to our society if we installed in public places. And hoping that this EBMT system can be efficiently used by everyone if we can release it as an open source. The instant messaging system (chat) has become an important medium for a huge number of people to communicate with each other. Even if the instant messaging technology is now so developed, there is a barrier to communication for people having different native-languages. We can overcome the barrier by integrating ”instant messaging” and ”machine translation” technologies Integration of this Malayalam to English translation system with a instant messaging system is one of the main applications of translation system. Also the system can be used in restaurants and hotels for translation of food menus. Also it helps in easy browsing of internet.

Journal Papers:

[1] Latha R Nair, David Peter S,Renjith P Ravindran, Design and Development of a Malayalam to English

[2] Translator- A Transfer Based Approach,*International journal of computational linguistics*, 2012, vol 3

[3] Harjinder Kaur,Dr. Vijay Laxmi A Web Based English to Punjabi MT System for News Headlines, *International Journal of Advanced Research in Computer Science and Software Engg* ,2013, Vol.03

[4] Unnikrishnan P,Antony P J, Soman K P, A Novel Approach for English to South Dravidian Language

[5] Statistical Machine Translation System , *International Journal on Computer Science and Engineering* Vol. 02, 2010

Theses:

[6] R. Harshawardan. , *Rule Based Machine Translation System For English to Malayalam Language*, Amrutha University,2011

[7] Nano Gough ,*Example-Based M achine Translation using the Marker Hypothesis* , 2005.

Proceedings Papers:

[8] Rahul.C, Dinunath.K, Remya Ravindran, K.P.Soman, Rule Based Reordering and Morphological Processing For Statistical English to Malayalam Translation,Proc. *International Conference on Advances in Computing, Control, and Telecommunication Technologies*, 2009, 458-460

[9] Sumam Mary Idicula,Anitha T Nair, Syntactic Based Machine Translation from English to lam,

[10] Proc.*International Conference on Data Science & Engineering*, 2012,.198-202

[11] Michael Paul, Eiichiro Sumita, and Seiichi Yamamoto. . Example-based rescoring of statistical ne translation output , Proc *First National Symposium on Modeling and Shallow Parsing of Indian Languages*, (MSPIL) ,2003

[12] Makoto Nagao A Framework Of a Mechanical Translation Between Japanese And English BY Analogy

[13] principle ,Proc. *Elsevier Science Publishedon artificial intellignece* , 1984.

[14] Rashmi Gangadharaiah and N. Balakrishnan Application of linguistic rules to generalized example based

[15] machine translation for indian languages , Proc *TheFirst National Symposium on Modeling and Shallow*

[16] *Parsing of Indian Languages*,2006.

[17] Ambati v and U.Rohini A Hybrid Approach to Example based Machine Translation for Indian Languages, Proc.*5th international conference on natural language processing* ,2007.