AN OVERVIEW OF COMPUTATIONAL LINGUISTICS AND MALAYALAM

Bijimol T.K.[1], Dr. John T. Abraham[2]

[1]*Research Scholar, Bharathiar University Coimbatore, Tamil Nadu (India)*

[2]*Asst. Professor, Dept. of Computer Science, Bharata Mata College, Thrikkakkara, Kerala (India)*

ABSTRACT

Computational linguistics is an interdisciplinary field related with the modeling of natural language from a computational viewpoint. As part of computational linguistics several Malayalam transliteration tools are available to the public. This tool allows you to type in Malayalam by its equivalent English characters. That means type the Malayalam words as it sound in English. A few machine translation systems are also developed in Malayalam for translating Malayalam to English and vice versa but these are not offered to the public. These works are based on Rule Based Machine Translation (RBMT) systems, Statistical Machine Translation systems, Example Based Machine Translation systems and Hybrid Machine Translation systems. This paper explains the efficiency of each system based on BLEU score or human evaluation. So many machine transliteration systems are available now but Malayalam translation systems are only in the beginning stage and needs to develop more.

Keywords*: *Computational Linguistics, Example Based System, Hybrid System, Machine Translation, Malayalam to English, Rule Based, Statistical, Transliteration.

I INTRODUCTION

Machine translation system is the part of computational linguistics which helps to convert one language to another. The importance of machine translation is increasing day to day because communication between different languages reduces the distance between two ends of the world. A wide range of research is going on machine translation area now and a fully automatic translation system is the aim of researchers. English is the international language so a lot of translation systems are concentrated on the conversion of English to local language and vice versa. Translation tools for other language pairs are also available.

Malayalam is the Dravidian language which has intricate nature and morphological richness. It is the mother tongue of Kerala. In order to communicate from Malayalam to English it is necessary to use a translation tool. English is the most popular language in the world but most of the people in Kerala are not so familiar with English. For the easy communication it is required to use the translation tools for converting Malayalam to English language in the applications like instant message systems, communication systems etc. On the other hand, many research people concentrated on English to Indian language translations. They are interested to review foreign websites, articles, and manuals. English to Malayalam translation systems satisfy the needs of these people. Machine translation is a good approach for localization of languages and it is a difficult job for translating morphologically rich and syntactically different languages.

II MACHINE TRANSLATION METHODS

There are four machine translation methods available now. Following are the brief explanation of these methods:

2.1 Rule Based Machine Translation (RBMT)

It selects rules from dictionaries and grammars based on the linguist information about source target language pairs [1]. Source language structure is directly converted into target language structures in this method. The advantages of RBMT are predictability and easy customization. At the same time lack of good dictionaries, expensive dictionary building and ambiguity in the rules are considered as its drawbacks. Eg. UNITRAN (UNIversal TRANslator)

2.2 Statistical Machine Translation (SMT)

SMT concept is coming from information theory and it uses the statistical models in order to generate the output. There is no customization work needed because translation tool learns methods from statistical analysis of bilingual corpora [2]. It is less expensive than RBMT and it has better resource usage. Corpus is the basis of this method but its creation is expensive with limited resources [3]. SMT does not work well with languages that have different word orders and it is not possible to predict the result in SMT. Eg: n-gram based SMT

2.3 Example Based Machine Translation (EBMT)

Examples are the basis of translation in EBMT system. It uses examples from the corpus to translate similar types of sentences [4]. It is possible to generate output quickly in this method because it uses examples to train the system. If the data set is small it works well but if deep linguistic analysis is required it act as inefficient. Eg. PanEBMT

2.4 Hybrid Machine Translation (HMT)

It combines the advantages of Rule Based systems and Statistical Machine Translation systems to translate source language to destination language [5]. HMT confers the better output comparing to other methods. It needs less post-editing of destination language. Accuracy of HMT is high and it provides fast and quality output. HMT uses the properties of RBMT and SMT so it combines the drawbacks of different approaches. Eg. DFKI-LT

III GLOBAL TRANSLATION SYSTEMS

The rapid growth of machine translation system leads to the development of commercial and non commercial translation systems. Commercial translation companies concentrated to develop quality products for the market [6]. It increases the competition and a good number of translation tools came into existence. Babylon, @prompt, Systran etc are some of the commercial machine translation software systems that are popular in the world. The study of current translation systems lead to the development of a website (http://translation-softwarereview.toptenreviews.com) which finds out top ten commercial translation tools worldwide. TABLE-1 shows the top ten profit-making machine translation tools which obtained from the above site:

Table 1. Top ten machine translation tools in the world

Position	Machine Translation Software
1	Babylon
2	Power Translator
3	Promt
4	WhiteSmoke
5	Translate Personal
6	Promt Personal
7	Translution
8	LingvoSoft Translator
9	IdiomaX
10	Ace Translator

IV COMPUTATIONAL LINGUISTICS AND MALAYALAM

4.1 Centre for Linguistic Computing Keralam

This is a joint undertaking of Computational Linguistic Team @ C-DIT, Kerala State IT Mission and Dept: of Linguistics, University of Kerala for research [7]. The use of information technology demands the need of local language to communicate to local people. For achieving this Computer Linguistic Team decided to develop software which helps the people in Kerala to understand and use information technology features in Malayalam. Their first product NILA was launched in 2004 that provides the facility to use e-governance and socialize the government programs through Malayalam. Next product Kaveri was launched in 2006. It assists text processing in Malayalam. As part of text processing it also provides transliteration system of major South Indian Languages and English-Malayalam translation help. Periyar is the next product (2010) and Pamba and Kabani are the incoming products of this team.

4.2 Malayalam transliteration tools

Several Malayalam transliteration tools are available in the websites. It is not translation tools but it allows you to type in Malayalam by its equivalent English characters. That means type the Malayalam words as it sound in English. These are also called as phonetic converters. Keraleeyam is one popular phonetic converter that is used by many users to type in Malayalam. TABLE-2 shows some phonetic converters available in the internet.

Table 2. Malayalam transliteration tools

Sl.No.	Malayalam Transliteration tools
1	Keraleeyam
2	Varamozhi
3	Changathi
4	Easy Malayalam
5	TamilCube
6	Aksharangal
7	English To Malayalam Translator
8	Yahoo! Transliteration
9	Malayalam Typing Software
10	QUILLPAD Writer
11	Softoni malayalam typeing

12	HariSree Malayalam Software Pack
13	MyMalayalam.com
14	LIPIKAAR
15	NHM Writer

4.3 Machine Translations Works in Malayalam

There are some translation systems for converting Malayalam sentences into its equivalent English translation and English to Malayalam translations. But the software implementations for the translations are not available till now. This section discusses the available studies on machine translations related to Malayalam.

4.3.1 English to Malayalam Translations

There is no tool available for English to Malayalam translation but various researches are progressing in this line.

R. Rajan et al. [8] have used the RBMT method in order to convert English sentences to Malayalam. It uses bilingual dictionaries and rules and is based on the parts of speech tag and dependency information from the parser. Transfer link rule and morphological rules are used here. English Malayalam bilingual dictionary is dictionary used in this work.

Aneena George [9] proposed an SMT system based approach for English to Malayalam conversion. In order to build an SMT it uses the probabilistic model. Parallel corpus is aligned by Berkeley word aligner and Hidden Markov model is used for the development of training and evaluation. Language model calculates the probability of destination language sentences and translation model computes the probability of target language sentences given the source language sentence with the help of Baum Welch training algorithm. 50 English and Malayalam sentences are included in the parallel corpus for train the system. The quality of the output in this system depends upon the size and quality of the corpus used for training.

Mary Priya Sebastian et al. [10] have also proposed an SMT system based approach which utilizes the bilingual English Malayalam corpus and monolingual Malayalam corpus in the training phase. Techniques to improve the alignment model, removing insignificant alignments, suffix separation from Malayalam corpus and stop word elimination from the bilingual corpus are the methods used for improving the efficiency of the system. Quality of output in this system is measured using BLEU score. Sentences in Training set has BLEU score 74 and Unseen sentences have 43.

Nithya B. et al. [11] have proposed an approach based on hybrid translation which is the combination of rule based and statistical machine translations. In order to perform hybrid English to Malayalam translation, statistical machine translator and translation memory caches are used. A statistical machine translator uses the machine learning techniques on the corpus. Translation memory caches are used for eliminating the redundant translations. This system is evaluated by BLUE score and human evaluation. This system proves 69.33 BLUE score and manual evaluation shows the accuracy of 75.3%.

J. Sangeetha et al. [12] have proposed a hybrid system based approach which translates English to Indian languages such as Tamil, Malayalam and Hindi. This system is the combination of statistical and ruled based methods. Rule based system builds rules which help to re-order the syntactic structures of source language. Context Free Grammars is applied for the generation of language structures and statistical technique is used to correct the errors in the translated text. Simplifying and segmenting are applied in source language for improving the quality of machine translation. This work is implemented with BLEU score 79.23.

4.3.2 Malayalam to English Translations

There are three works available under the category of Malayalam to English translations. These are based on RBMT, EBMT and Hybrid translation methods.

A transfer based RBMT approach is proposed by *Latha R Nair et al.* [13] for Malayalam to English translations. This system includes preprocessor for dividing the compound words, a syntactic structure transfer module, bilingual dictionary and morphological parser (context disambiguation and chunking). Rules for Malayalam morphology and rules for syntactic structure transfer are two rules used in this system. Artificial intelligence techniques are used here and it is possible to build translation system for other language pairs easily. After the implementation about 20% of output shows the exact translation of input and remaining were meaningful but small limitations due to some reasons.

Another Malayalam to English translation system is proposed by *Anju E.S et al.* [14] using EBMT method that has three steps such as example acquisition, matching and recombination. This is evaluated by human experts based on the perfection of the translated output. System works well for the simple sentences with 75% quality in translation and remaining have reordering problems. In order to improve the performance it is recommended to use large aligned corpus and more reordering rules.

Rajesh. K. S et al. [15] have proposed a Hybrid approach which is the combination of word-aligned parallel corpus based and dictionary lookup methods. First three IBM models and Expectation Maximization (EM) algorithm are the basis of corpus and bilingual Malayalam-English Dictionary is used for dictionary lookup approach. Malayalam-English corpus has 950 sentence pairs and 255 sentence pairs were used for testing. This approach gives 91% precision for translated output.

V RESEARCH FINDINGS

Translation systems in Malayalam are in developing stage. There are many profitable and non profitable systems globally available but the Malayalam translation tools require long way to reach to this perfection. A lot of open source transliteration and free systems are available to the public but it provides only Malayalam sentences when typing equivalent English sounds. There is some translations systems are developed but as a tool it is not available to public and its efficiency is questionable. So it can conclude that the Malayalam translation tools are in developing stage compared to global translation tools and it requires lots of research to reach to perfection.

VI CONCLUSION

Machine translation is the process of translating one natural language to another with the help of computer system. Numerous translation systems are available in the market today. Translation systems in Malayalam are in developing stage and computer linguistic team of CDIT contributes some text processing systems including translation aid to the area of machine translation. Several machine transliteration tools are available today but that does not translate Malayalam to other languages or vice versa. Some Malayalam machine translation systems were developed but its implementation is not available to the people. It is an acceptable fact that translation systems in Malayalam are in emerging stage and requires lots of research in this regard.

REFERENCE

[1] Marta R. Costa-Jussa`,Mireia Farrus, Jos´e B. Marino˜ , Jos´e A. R. Fonollosa, Study And Comparison Of Rule-Based And Statistical Catalan-Spanish Machine Translation Systems, Computing and Informatics, Vol. 31, 2012, pp. 245–270

[2] S. Tripathi, J. K. Sarjgek, Approaches to machine translation, Annals of Library and Information Studies, Vol. 57, 2010, pp. 388-393

[3] C. Dove, O. Loskutova, and R. Fuente, What's Your Pick: RbMT, SMT or Hybrid?, The Tenth Biennial Conference of the Association for Machine Translation in the Americas, 2012

[4] E. Sumita, H. Iida, Experiments and Prospects of Example-based Machine Translation, ACL Anthology Reference Corpus, 2010, pp. 185-192

[5] Sabine Hunsicker, Chen Yu, Christian Federmann, Machine Learning for Hybrid Machine Translation, Proceedings of the 7th Workshop on Statistical Machine Translation, , Montreal, Canada, June 7-8, 2012, pp. 312–316

[6] Stephen Hampshire , Carmen Porta Salvia, Translation and the Internet: Evaluating the Quality of Free Online Machine Translators, Quaderns. Rev. trad. 17, 2010, pp. 197-209

[7] Computer linguistic team CDIT, Malayalam Computing Solutions Provider for the Government of Kerala and various other agencies in Kerala,
available at: http://www.cdit.org/computionallinguistic.htm

[8] R. Rajan, R. Sivan, R. Ravindran, K.P. Soman, Rule Based Machine Translation from English to Malayalam, International Conference on Advances in Computing, Control, & Telecommunication Technologies, 2009.

[9] Aneena George, English To Malayalam Statistical Machine Translation System, Article of International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 7, 2013.

[10] Mary Priya Sebastian, Sheena Kurian KB, G. Santhosh Kumara, A framework for translating English text into Malayalam using statistical models , 2nd International Conference on Communication, Computing & Security, 2011.

[11] Nithya B, Shibily Joseph, A Hybrid Approach to English to Malayalam Machine Translation, International Journal of Computer Applications (0975 – 8887) Volume 81 – No.8, 2013.

[12] J. Sangeetha, S. Jothilakshmi ,R.N.Devendra Kumar, An Efficient Machine Translation System for English to Indian Languages Using Hybrid Mechanism, International Journal of Engineering and Technology (IJET).

[13] Latha R Nair, David Peter S, Renjith P Ravindran, Design and Development of a Malayalam to English Translator- A Transfer Based Approach, International Journal of Computational Linguistics (IJCL), Volume (3) : Issue (1) : 2012.

[14] Anju E S, Manoj Kumar K V, Malayalam To English Machine Translation: An EBMT System, IOSR Journal of Engineering (IOSRJEN), www.iosrjen.org, Vol. 04, Issue 01, 2014.

[15] Rajesh. K. S, Veena A Kumar & CH. Dayakar Reddy, Building a Bilingual Corpus based on Hybrid Approach for Malayalam-English Machine Translation, Special Issue of International Journal of Computer Science & Informatics (IJCSI), Vol.- II, Issue-1, p. 2