# English To Malayalam Statistical Machine Translation System

Aneena George
*Adi Shankara College of Engineering and technology*

## Abstract

*Machine Translation is an important part of Natural Language Processing. It refers to a machine to convert from one natural language to another. Statistical Machine Translation is a part of Machine Translation that strives to use machine learning paradigm towards translating text. Statistical Machine Translation contains a Language Model (LM), Translation Model (TM) and a Decoder. Statistical Machine Translation is an approach to translating source to target language. In our approach to building SMT we use a probabilistic model. Here Bayesian network model as Hidden Markov Model (HMM) is used for designing SMT.Berkeley word aligner is used for aligning the parallel corpus. In this thesis, English to Malayalam Statistical Machine Translation system has been developed. The development of Training and Evaluation is done by using hidden markov model.LM computes the probability of target language sentences. TM computes the probability of target sentences given the source sentence by using training algorithm Baum Welch algorithm and the Evaluation maximizes the probability of translated text of target language. A parallel corpus of 50 simple sentences in English and Malayalam has been used in training of the system.*

## 1. Introduction

The technology is reaching new heights, right from conception of ideas up to the practical implementation. It is important, that equal emphasis is put to remove the language divide which causes communication gap among different sections of societies. Natural Language Processing (NLP) is the field that strives to fill this gap. Machine Translation (MT) mainly deals with transformation of one language to another. Machine Translation (MT) is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another [1]. At its basic level, MT performs simple substitution of words in one natural language for words in another. Current machine translation software often allows for customization by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions. This technique is effective in domains where formal or formulaic language is used. It follows that machine translation of legal documents more readily produces usable output than conversation or less standardized text [1].

Machine Translation system are needed to translate literary works which from any language into native languages. The literary work is fed to the MT system and translation is done. Such MT systems can break the language barriers by making available work rich sources of literature available to people across the world.

MT also overcomes the technological barriers. Most of the information available is in English which is understood by only 3% of the population [2]. This has led to digital divide in which only small section of society can understand the content presented in digital format. MT can help in this regard to overcome the digital divide.

Statistical Machine Translation (SMT) is a probabilistic framework for translating text from one language to another, based on parallel corpus. [3]The first ideas of statistical machine translation were introduced by Warren Weaver in 1949, including the ideas of applying Claude Shannon's information theory. Statistical machine translation was re-introduced by researchers at IBM's Thomas J in 1991, Watson Research Centre and has contributed to the significant resurgence in interest in machine translation in recent years. The idea behind statistical machine translation comes from Information Theory. A document is translated according to the probability distribution that a string in the target language (for example, MALAYALAM) is the translation of a string in the source language (for example, ENGLISH).

### 1.1 Problem Statement

With each passing day the world is becoming a global village. There are hundreds of languages being spoken across the world. The official languages of different states and nations are also different according to their cultural and geographical differences.

Most of the content available in digital format is in English language. The content shown in English must be presented in a language which can be understood by

the intended audience. There is large section of population at both national and state level who cannot comprehend English language. It has brought about language barrier in the side lines of digital age. Machine Translation (MT), can overcome this barrier. In this thesis, a proposed Statistical Based Machine Translation system for translating English text to Malayalam language has been proposed. English is the source language and the Malayalam is the target language.

The Problem defined here is how to translate English text to Malayalam text by using statistical approach with Hidden Markov Model (HMM) as a concept of proof.

## 1.2 Existing MT System

There are following MT systems that have been developed for various natural language pair.

### 1.2.1 Systran

Systran is a rule based Machine Translation System developed by the company named Systran. It was founded by Dr. Peter Toma in 1968. It offers translation in text from and into 52 languages. It provides technology for Yahoo! Babel Fish and it was used by Google till 2007 [2]. In 2009 SYSTRAN extended its position as the industry's leading innovator by introducing the first hybrid machine translation engine.

### 1.2.2 Google Translate

Google Translate is service provided by Google to translate a section of text, or a webpage, into another language. The service limits the number of paragraphs, or range of technical terms, that will be translated [13]. Google translate is based on Statistical Machine Translation approach.

### 1.2.3 Bing Translator

Bing Translator is a service provided by Microsoft, which was known as Live Search Translator and Windows Live Translator. It is based on Statistical Machine Translation approach.

Four bilingual views are available:

· Side by side
· Top and bottom
· Original with hover translation
· Translation with hover original

## 1.3 Proposed System

The SMT system is based on the view that every sentence in a language has a possible translation in another language. A sentence can be translated from one language to another in many possible ways. Statistical translation approaches take the view that sentence in the target language is a possible translation of the input sentences [3].

The main intent of having a statistical based approach to translation is to give the end user the freedom from employing large translation teams to get the translation of texts. This is particularly important when the application is in like fields. For eg: if the intent is to translate children's books, the input should be in that area. Using the SMT is able to make a wise decision on what the input data would be.

The benefits of statistical machine translation over traditional paradigms are:

- Better use of resource
- There is a deal of natural language in machine-readable format.
- More natural translations
- A SMT would greatly increase the resource utilization (disk and cpu) as compared to the rule based system
- Decrease the dependency on language translations on a language expert.
- Higher accuracy provide for domain specific application like weather report, medical domine etc...
- SMT depends on size of corpus, type of corpus and domain of corpus
- Accuracy of SMT can improved by increasing the resources like parallel corpus and trained corpus
- In rule based system accuracy can improved by rule modification, it is a tedious task

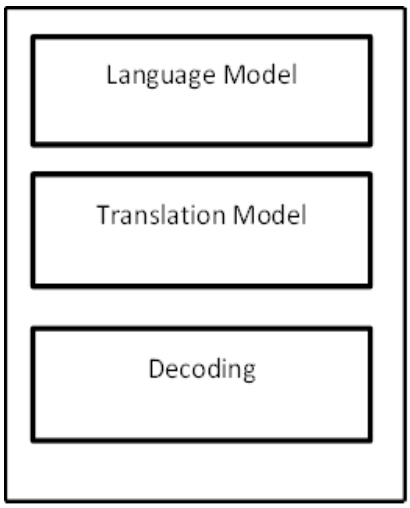


**Figure-1.Outline of statistical machine translation system**

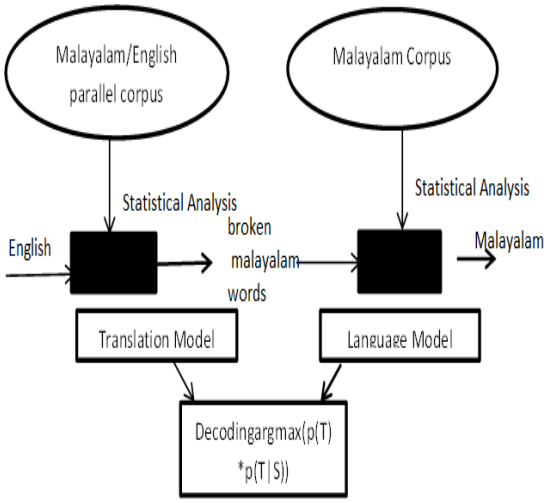


**Figure- 2. Working of SMT**

### 1.3.1 Language Model

A language model gives the probability of a sentence. The probability is computed using *n-gram* model. Language Model can be considered as computation of the probability of single word given all of the words that precede it in a sentence [4].

The goal of Statistical Machine Translation is to estimate the probability (likelihood) of a sentence. A sentence is decomposed into the product of conditional probability. By using chain rule, this is made possible as shown in 2.1. The probability of sentence (S) is broken down as the probability of individual words P (w).

P(s) = P(w1, w2, w3,....., wn)

=P (w1) P (w2|w1) P (w3|w1w2) P (w4|w1w2w3)…P (wn|w1w2…wn-1)) … (2.1)

In order to calculate sentence probability, it is required to calculate the probability of a word, given the sequence of word preceding it. An *n-gram* model simplifies the task by approximating the probability of a word given all the previous words. An *n*-gram of size 1 is referred to as a *unigram*; size 2 is a *bigram* (or, less commonly, a *diagram*); size 3 is a *trigram*; size 4 is a *four-gram* and size 5 or more is simply called an *n-gram*.

Consider the following training set of data:

There was a King
He was a strong King.
King ruled most parts of the world.

Training set of data for LM:

Probabilities for bigram model are as shown below:

P (there/<s>) = 0.67 P (was/there) = 0.4 P (king/a) = 1.0 P (a/<s>) =0.30 … (2.2)

P (was/he) = 1.0 P (a/was) = 0.5 P (strong/a) = 0.2 P (king/strong) =0.23 ... (2.3)

P (ruled/he) = 1.0 P (most/rules) = 1.0 P (the/of) = 1.0 ... (2.4)

P (world/the) =0.30 P (ruled|king) =0.30 ... (2.5)

The probability of a sentence: 'A strong king ruled the world', can be computed as

Follows:

P (a/<s>)* P (strong/a)* P (king|strong)*P (ruled|king)*P (the/ruled)*P (world|the)

=0.30*0.2*0.23*0.30*0.28*.0.30

=0.00071

### 1.3.2 Translation Model

The Translation Model helps to compute the conditional probability *P (T|S)*. It is trained from parallel corpus of target-source pairs. As no corpus is large enough to allow the computation translation model probabilities at sentence level, so the process is broken down into smaller units, *e.g.*, words or phrases and their probabilities learn [4]. The target translation of source sentence is thought of as being generated from source word by word. For example, using the notation *(T/S)* to represent an input sentence *S* and its translation *T*. Using this notation, sentence is translated as given in the below sentence.

(Patti poothottathil kidkkunnu | dog slept in the garden)

(പട്ടി പൂതോട്ടത്തിൽ കിടക്കുന്നു | dog slept in the garden)... (2.7)

One possible alignment for the pair of sentences can be represented as given in 2.8:

(പട്ടി പൂതോട്ടത്തിൽ കിടക്കുന്നു | dog (1) slept (3) in (null) the (null) garden (2))... (2.8)

A number of alignments are possible. For simplicity, word by word alignment of Translation model is considered. The above set of alignment is denoted as *A(S, T)*. IfLength of target is *l* and that of source is *m* than there are *lm* different alignments arePossible and all connection for each target position are equally likely, therefore orderOf words in T and S does not affect *P (T|S)* and likelihood of *(T|S)* can be defined in Terms of the conditional probability *P (T, a/S)* as,

*P (S|T)* = sum *P(S, a/T)* ... (2.9)

The sum is over the elements of alignment set, *A(S, T)*. English word has only exactly one connection for the alignment,

P(പട്ടി പൂതോട്ടത്തിൽ കിടക്കുന്നു | dog slept in the garden), can be computed by multiplying the translation probabilities T(പട്ടി |dog(1)),

T(പൂതോട്ടത്തി | garden(6)), T(null|in(3)), T(null|the(4)), and T(കിടക്കുന്നു | slept(2)).

**1.3.3 Decoder**

This phase of SMT maximizes the probability of translated text. The words are chosen Which have maximum like hood of being the translated translation [5]Search for sentence T is performed that maximizes *P (S|T) i.e.*

Pr (S, T) = argmax P (T) P (S|T)

**1.4 Objective**

The objectives of thesis are as under:

1. To understand the Bayesian network model as Hidden Markov Model for SMT
2. To understand the Berkeley word aligner
3. To understand the Language Model (LM), Translation Model (TM) of SMT.
4. To create a LM for Malayalam with use of Ngram model.
5. To generate Malayalam and English parallel corpus for training the system
6. Baum Welch algorithm is used for Training the corpus

The objective is to create a STATISTICAL MACHINE TRANSLATION (SMT) system for English to Malayalam as a concept of proof.

## 2 Materials and Methods

**2.1 System Requirements**

1. Intel i7 processor
2. Mac OS with Malayalam font installed
3. Java 1.6 or above

## 2.2 SMT Analysis

**2.2.1 Development of Corpus**

Statistical Machine Translation system makes use of a parallel corpus of source and target language pairs. This parallel corpus is necessary requirement before undertaking training in Statistical Machine Translation. The proposed system has used parallel corpus of English and Malayalam sentences. A parallel corpus of more than 100 sentences has been developed from which consist of small sentences and the life history of freedom fighters with reference to their trail in courts.For example a list of parallel corpus is given below.

**Table1: English and Malayalam parallel corpus**

| Bitext.e | Bitext.f |
|---|---|
| I am aneena | ഞാൻ അനീന ആകുന്നു |
| I am anju | ഞാൻ അഞ്ജു ആണ് |
| I am arun | ഞാൻ അരുൺ ആണ് |
| I am a good boy | ഞാൻ ഒരു നല്ല കുട്ടി ആണ് |
| I am a bad boy | ഞാൻഒരു ചീത്ത കുട്ടി ആണ് |
| I am a boy | ഞാൻ ഒരു ആൺകുട്ടി ആണ് |
| I am a girl | ഞാൻ ഒരു പെൺകുട്ടി ആണ് |
| My name is aneena | എൻറെ പേര് അനീന ആകുന്നു |
| My name is arun | എൻറെ പേര് അരുൺ ആകുന്നു |

**2.2.2 Berkeley Word Aligner**

The Berkeley Word Aligner is a statistical machine translation tool that automatically aligns words in a parallel corpus.

**2.2.3 Hidden Markov Model(HMM)**

***Markov models***

Markov models are used to model sequences of events (or observations) that occur one after another.The easiest sequences to model are deterministic, where one specific observation always follows another,Example: changes in traffic lights (green to yellow to red).In a nondeterministic Markov model, an event might be followed by one of several subsequent events, each with different probability

– Daily changes in the weather (sunny, cloudy, rainy)
— Sequences of words in sentences
– Sequences of phonemes in spoken words

A Markov model consists of a finite set of states together with probabilities for transitioning from state to state. Consider a Markov model of the various pronunciations of "tomato":

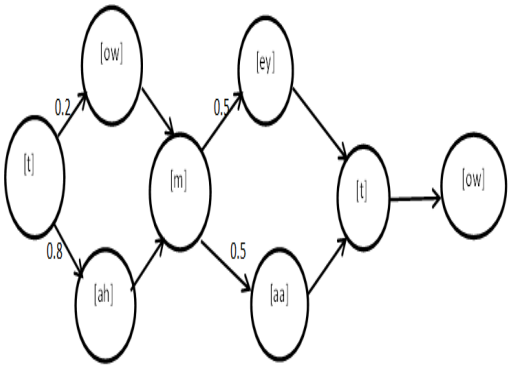


**Figure-3 State transition diagram of tomato**

The probability of each path is the product of the probabilities on the arcs that create the path

P ([towmeytow]) = P ([towmaatow]) = 0.1

P ([tahmeytow]) = P ([tahmaatow]) = 0.4

Transition from state i to j is given by the probability aij = P (sj|si)

The state transition probabilities of states n model is represented by an n × n matrix

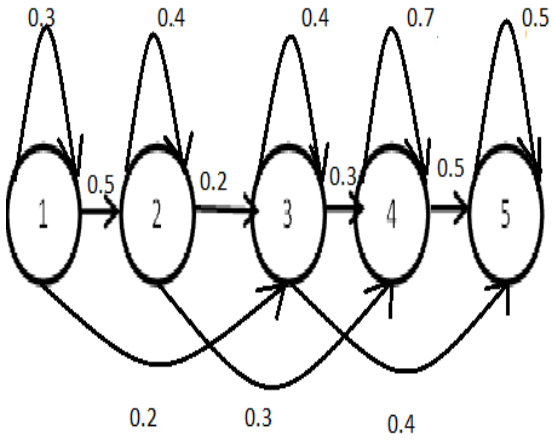


**Figure- 4 State transition diagram**

$$A_{ij} = \begin{bmatrix} 0.3 & 0.5 & 0.2 & 0 & 0 \\ 0 & 0.4 & 0.3 & 0.3 & 0 \\ 0 & 0 & 0.4 & 0.3 & 0.4 \\ 0 & 0 & 0 & 0.7 & 0.5 \\ 0 & 0 & 0 & 0 & 0.5 \end{bmatrix}$$

In an ordinary Markov model the output (sequence of observations) is simply the sequence of states visited: [towmeytow]

There are also Hidden Markov models (HMMs), where the notions of observation and state are separated

– States are not represent observations directly

– Different states produces different outputs

– The output is not the set of states visited

An HMM is specified by a set of states *s*, a set of transition probabilities *a*, and a set of observation *b*. *bj(ot)* is the probability of emitting symbol *ot* when state *sj* is entered at time t.[16]

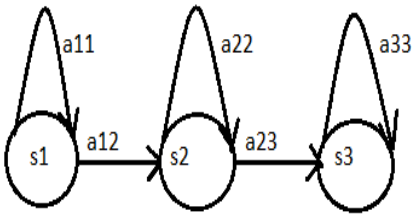


**Figure- 5 State transition diagram of HMM**

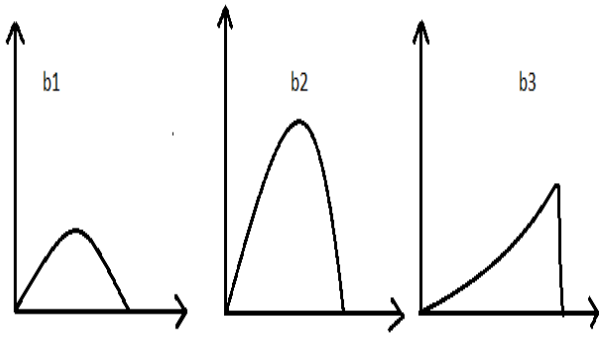


**Figure- 6 Sequence of emissions**

*HMM Definition*

HMM can be defined as a set of 5 elements

1. S representing the set of states s1,s2,s3,...si, sn
2. V, the output symbols/ observation/emissions v1,v2,v3,...,vm
3. p(i), the probability of being in state qi at time t=0 .That is initially which state will the model will be in. So there will be probabilities associated with this. π gives this matrix.
4. A=transition probabilities =aij where aij= Probability of going to state qj at time t+1 given that the model is in state qi at time t.
5. B=output probabilities, bj(k), where bj(k) = Probability of producing symbol vk from state qj at time t.

*HMM Assumptions*

(a)Markov assumption next state is dependent only upon the current state.

(b) Stationery assumption state transition probability is independent of the actual time at which the transition takes place.

(c) Output Independence assumption current output is statistically independent of the previous outputs (observations). [8]

***HMM Activities***

Three basic activities that must be solved for the HMM model to be useful in real world applications are

- Evaluation
- Decoding
- Learning (Training)

**Evaluation**

Application 1

Given a model and a sequence of observations. We need to find out the probability that the observations are generated by the model. Basically we score how well a given model matches a given observation sequence.

Application 2

If we have a number of HMMs describing deferent systems, and a sequence of observations. We want to know that which HMM is most probably generated the given sequence.

**Decoding**

To find the hidden states that generated the observed output. Thus we try to uncover the hidden part of the model i.e. correct state sequence. Viterbi algorithm to determine the most probable sequence of hidden states given a sequence of observations and a HMM.

**Learning**

The third, and the most tricky and challenging problem associated with HMMs is determining the model parameters most likely to have generated a sequence of observations that is - to fit the most probable HMM; that most probably describes what is seen.In this **Baum Welch** algorithm is used for training.

**Baum Welch algorithm [Forward and Backward]**

Finding the transition probability matrix A by using the forward and backward algorithms.

**1.Initialize the sequence of elements{A, B, Pi}For Learning**

A-Transition Probability
B-Emission Probability
Pi-Initial state probability

*1.1 Initialize- A*

States are hidden that is taken from obsevations

```
For (int i = 0; i < states; i++)
For (int j = i; j < states; j++)
   A[i] [j] = 1.0 / (states - i)
```

*1.2 Initialize-B*

```
For (int i = 0; i < States; i++)
For (int j = 0; j < outputSize; j++)
     B[i] [j] = 1.0 / output Size
```

*1.3 Initialize –Pi*

Pi=0

**2. Forward Algorithm**

*2.1. Initialization*

T=observations.length;

fwd [0] [i] = pi[i] * B[i] [observations [0]];

*2.2 Induction*

$$fwd[t][i] = \sum_{t=1}^{T-1} \sum_{i=0}^{states-1} \sum_{j=0}^{states-1} (fwd[t-1][j] * A[j][i]) * P$$

$$C[t] = \sum_{t=1}^{N-1} \sum_{i=0}^{states-1} fwd[t][i];$$

**3 Backward Algorithm**

*3.1. Initialization*

```
for (int i = 0; i < States; i++)
   bwd [T - 1] [i] = 1.0 /c [T - 1];
```

*3.2. Induction*

$$bwd[t][i] = \sum_{i=0}^{states-1} \sum_{j=0}^{states-1} (A[i][j] * B[j][observations[t+1]] * bwd[t+1][j])/c[t]$$ for $t = T-2$

## 4 Design of SMT

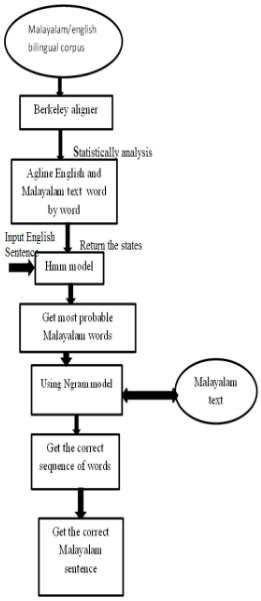


**Figure- 7 Working of SMT**

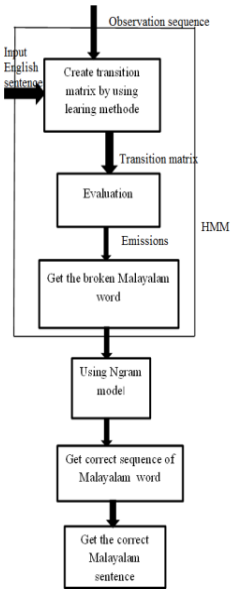


**Figure -8 Working of HMM**

### 3.2 Working of SMT

For the SMT to work we need to use a probabilistic model. The model used here is HMM (Hidden Markov Models). Markov chains have always been a model for learning algorithms. Here states are predicted probabilistically based on observations. HMM is a specialized form of Markov chains for which the hidden is the state.

The input to the model is the observation, here the observation are the states which are being modelled. When a state moves this gives us transition. The more the state moves the lesser is the accuracy by which we can predict outcome (emission)

In the above paragraph we have laid down the basic principles of HMM. The same is used to model further the problem at hand which is the translation of English to Malayalam States are the Malayalam translations for an English word and the transitions are the occurrences in which the English word aligns to different words in Malayalam.

The first Malayalam word is state 1 etc. So an ideal observation would be {0, 1,1,1,1...1} n times. This means that we get the same state for the English word. However this is not always the case. We get{0,1,1,2,1,2,1,1,2,2,..}. So to increase the EM (expectation maximization) of the model what we need is a stable state distribution which means that even if we don't achieve {0, 1, 1, 1,1} we should be able to get a distribution which when compared against {0,1} gives a demarkating probability.

For this in this project we have selected Berkeley aligner which has been developed by UC Berkeley. This is currently one of the best aligners. This helps us achieve the stable state distribution. Once we have alignments the states are fed into the HMM which uses Baum Welch algorithm to come up with the transition matrix. Once we have the transition matrix the emissions probability for a sequence of states can be calculated using evaluation.

Once we have the emissions we need a language model which arranges the words in proper order and follow up the words with missing words which complete the meaning. A Turing machine is the best for this. Here the turing machine is implemented by n-gram model (currently we use bigrams).The culmination of all this is the final Malayalam translated sentence.

## 4 Result

**Input:**My friend is a bad kid

**Output** എൻറെ സുഹൃത്തു ഒരു ചീത്ത കുട്ടി ആണ്

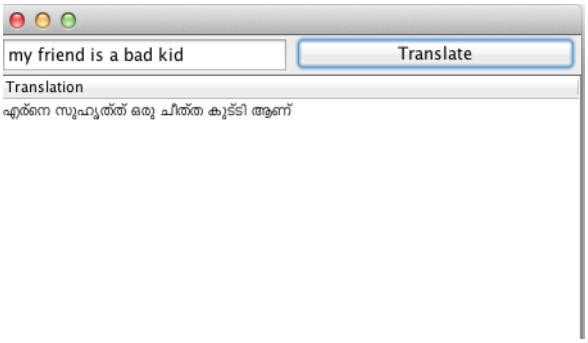

**Figure-9 Screen shot for translation**

## 5 Conclusion

In this thesis, English to Malayalam SMT system has been developed. The SMT is a part of Corpus basedMT system which requires parallel corpus before undertaking translation. A parallel corpus of 50 English and Malayalam sentences was used to train the system. The SMT system developed accepts English sentences as input and generates corresponding translation in Malayalam. The quality of the translated text can be depends upon the size of the corpus and the quality of the corpus.

## 6 Future Work

There can be following future directions for English to Malayalam SMT system.

- The work can be extended to include bidirectional translation between English and Malayalam
- The corpus can be pre-processed to change its clause structure for improving the quality of translation.
- The translated text can be reordered and processed to overcome grammatical mistakes which will be part of post-processing. This will improve score of human evaluation

## 7 Acknowledgement

I thank my project guide Mr Jayan.V , Mr Bhadran V.K. CDAC-Trivandrum and Abraham Varghese Adi Shankara Institute Of Engineering and Technology,Kalady for their valuable guidance,monitoring and support in the research area of Natural Language Processing throughtout the course of this work. I also wish to express my gratitude and obligation to C-DAC for giving their resources to complete this work.

## 8 References

[1] Tanveer Siddiqui and U.S. Tiwary, "Natural language Processing and Information Retrieval", New Delhi, Oxford Press, 2008

[2] D. D. Rao, "Machine Translation A Gentle Introduction", RESONANCE,July 1998.

[3] Amittai E. Axel. Factored language models for statistical machine translation.

[4] Statistical Machine Translation, Ananthakrishnan Ramanathan

[5] english to hindi statistical machine translation system, Nakul Sharma

[6] machine translation system in indian perspectives Sanjay Kumar Dwivedi and Pramod Premdas Sukhadeve Department of Computer Science, Babasaheb Bhimraom Ambedkar University, Lucknow, India.

[7]"Statisticalmachinetranslation",[Online].Available,http://en.wikipedia.org/wiki/Statistical_machinetranslation

[8] Dayne Freitag Andrew McCallum and Fernando PereiraHMM. Maximum entropy markov models for information extraction and segmentation.

[9]Hindi to Punjabi Machine Translation SystemVishal Goyal,Gurpreet Singh Lehal

[10] Machine Translation System in Indian Perspectives Sanjay Kumar Dwivedi and Pramod Premdas Sukhadeve Department of Computer Science, Babasaheb Bhimrao Ambedkar University,

[11]Survey of Indian Machine Translation Systems 1Sitender, 2Seema Bawa1,2Dept. of CSE, Thapar University, Patiala, Punjab, India V

[12]D. D. Rao, "Machine Translation A Gentle Introduction", RESONANCE,July 1998.

[13] Och, Franz Josef, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Association for Computational Linguistics, pp. 858-867,June2007,[Online]Available:http://www.translate.google.com.,http://translate.google.com/about/intl/en_ALL/

[14] "MachineTranslation",[Online].Available,http://faculty.ksu.edu.sa/homiedan/Publications/Machine%20Translation.pdf

[15] Vijayanand Kommaluri, Sirajul Islam Choudhury,Pranab Ratna,"VAASAANUBAADA"AutomaticMachineTranslation of Bilingual Bengali-Assamese News Texts

[16] HIDDEN MARKOV MODELS,John Fry San Jos´e State University