International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 1, January 2017

ISSN: 2394-4404

# Enhancing Machine Translation Using Graph Approach

**Komal S.Tilekar**

**Prof. Hema V Kumbhar**

Department of Computer Engineering,
Padmabhushan VasantDada Patil Institute of Technology,
Bavdhan, Pune

***Abstract:** In the present world, the things get change and the use of automatic machines are increasingly popular, the intelligent terminals (such as mobiles phones devices, personal computers, laptops and smart phones). The different methodologies or tools that are available for machine translation are Google Translate, Bing Translator, IBM, Moses, etc. Many of the existing system perform not up-to-the mark due to greater time complexity or other constraints. So this leaves a space for some contributions. This paper will focus on the problem to make the extremely interactive user-friendly statistical machine translation system which will compute the linguistic system which can be the toolkit of the smart devices work as a human translator. This method can change domain consultants to modify/add new translation rules so as to boost translation quality. A statistical machine translation (SMT) model that uses hierarchical phrases (HPBT) that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a parallel text without any syntactic annotations. Thus, system will performs significantly better than the Alignment Template System, a state-of-the-art phrase- based system.*

***Keywords**: Lexical Parser, Machine Translation, Rule Based Translation, POS tagging*

## I. INTRODUCTION

In this research work we are converting the simple English affirmative sentences to Hindi sentences. This is basically a machine translation. We have chosen the transfer based approach which is the thin line between the semantic and the direct approach. For that we have designed the parser which helps us to map the English sentence binding to the rules and then getting converted into target Language. English to Hindi language Translator (EHLT) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) language.

EHLT systems convert information from computer databases into readable human language. Natural systems convert samples of human language into more formal representations such as parse trees or first-order logic structures that are easier for computer programs to manipulate. The most explanatory method for presenting what actually happens within a NaturalLanguage Processing system is by means of the 'levels of language' approach. This is also referred to as the synchronic model of language and is distinguished from the earlier sequential model, which hypothesizes that the levels of human language processing follow one another in a strictly sequential manner. The morphological level deals with the componential nature of words, which are composed of morphemes – the smallest units of meaning.

At the lexical level, humans, as well as NLP systems, interpret the meaning of individual words. Several types of processing contribute to word-level understanding – the first of these being assignment of a single part-of-speech tag to each word. In this processing, words that can function as more than one part-of-speech are assigned the most probable part-of speech tag based on the context in which they occur.

Additionally at the lexical level, those words that have only one possible sense or meaning can be replaced by a semantic representation of that meaning. Rule-Based Machine Translation (RBMT also known as "Knowledge-Based

Machine Translation"; "Classical Approach" of MT) is a general term that denotes machine translation systems based on linguistic information about source and target languages basically retrieved from (unilingual, bilingual or multilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively. Having input sentences (in some source language), an RBMT system generates them to output sentences (in some target language) on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages involved in a concrete translation task.

This paper is classified as below where section 2 discusses past works as Literature Survey. Section 3 reveals the detailing of our proposed methodology. The evolution of our methodology is performed in section 4. Finally section 5 concludes this paper with some scope for future extensions.

## II. LITERATURE SURVEY

In statistical machine translation, we use the bilingual corpus in order to improve the quality of translation. [1] discusses about its viewpoints regarding the use of corpus. We can collect more and more bilingual corpus, such as extracting the sentence pairs from the comparable corpus. But if we consume larger corpus, more resources will be used. The solution for this problem can be that we can select only a part of corpus to train the translation model that will reduce the time and space complexity while building machine translation system and also will not decrease the translation quality.

In [2], they combine the phrase-based models and group of reconstructing rules on English sentence for producing reordering English sentence. They also incorporate morphological information by using a Morph Analyzer. Since morphological and parsing tools are not much widely available for Indian languages, an approach like this, which minimizes the use of such tools for the target language, would be quite handy.

A natural language word (aka, *surface form*) may be used in different contexts of sentences and thus represents different meaning. Morphological Analysis (MA) is a method used for decoding the meaning kept in every word and generates *root* (i.e., main meaning bearing unit of the word) of the word, the lexical category of the root, and associated grammatical features. As the words are having multiple grammatical functionalities when they are used in different contexts, they will have alternative morphological interpretations. It completely depends on the type of language that the average number morphological alternatives of a word will have.

There may be common vocabulary words of morphologically rich. The categorical pruning can be done with the Part-of-Speech tags of the words. Also, we can define rules to prune out some analyses having incompatibility with local dependencies. Bigram-based pruning approach is used for pruning alternatives having roots of same lexical class.

In [3], they have systematically pruned out the morphological analyses of the words which are incompatible with the context of the sentence. For this purpose, they have used extracted and used various syntactic information (like Part-Of-Speech, chunk properties) of the input sentences. Categorical pruning is done with the Part-of-Speech tags of the words. Rules are defined to prune out some analyses having incompatibility with local dependencies.

[4] paper tells us that they model the content of phrases, use the topic models, and estimate how much the source phrase is similar to the target phrase by each phrase pair. So the pruning of the irrelevant and low-quality phrase pairs from the phrase-table can be done by checking their content instead of just their number of occurrences in the training corpora, unlike the other common pruning methods. There may be many phrases that have low co-occurrences, but have the same meanings in the phrase-table. The methods which are just based on the counts and co-occurrences will prune these phrase pairs as noisy ones. Thus, this method helps us to prune over 50% of the phrase table without much loss in translation quality.

There is a concept of OAK Parser which is used to analyze the input English text to get the part of speech (POS) for each word in the text as a pre-translation process using the C# language. There are validation rules that can be applied in both the database design and the programming code in order to ensure the integrity of data. A major design goal of [5] system is that it will be used as a stand-alone tool, and can be very well integrated with a general machine translation system for English sentences But there were many shortcomings in the output of MT, due to either faulty analysis of the source language text or faulty generation of the target language text.

In [6], a technique is specifically developed for scenarios where bilingual corpora are provided. Furthermore, It is stated that the chance to understand that how the phrase-based extraction model has increased both the phrasal coverage and translation accuracy of the syntax-based model. So by following such approach, we are permitted to use the phrase translation technique into a statistical MT approach as well. This can be achieved by the help of Link Parser which is a sentence parser available in Link Grammar Formalism.

MT methodologies are broadly classified into two, namely rule based approach and corpus based approach. This research work discussed about only rule based approach. Vauquois triangle is the concept behind this. It basically tells to transform the source language into a target language. In this method, there is an increased amount of analysis, on both sides. The content of [7] mostly directs us to focus on syntax and capturing syntactic variance between English and Tamil. Synchronous Tree Adjoining Grammar (STAG) formalism can be used to represent both English and Tamil structures. Encouraging results thus, can be obtained even with very limited target resources. So there is a need to develop the statistical parser which can demonstrate the right parse from the erroneous parses. But the attribute of speed remains inferior.

In the process of machine translation, handling prepositions is the main issue. There are many various kinds of prepositions that are being used in English, which are translated into postpositions in Telugu. For selecting the appropriate postposition in Telugu, time, gender, place, context and many other features have an important part. There will be various situations when there will be different meanings of the same prepositions and the appropriate one should be selected based on the context. Through the proposed algorithm frequently used prepositions can be handled and translated perfectly.

The [8] states about writing effective set of rules and building a dictionary that is rich in words, so that the translation system can deal with any kind of sentences. Furthermore, it can also be extended to identify the phrases and idioms, other functional words and translate them in a better manner instead of direct translation.

## III. PROPOSED METHODOLOGY

Project can be design and implement with the following modules of development cycle. Research work implementation is been designed with layered architecture approach. Following are four layers

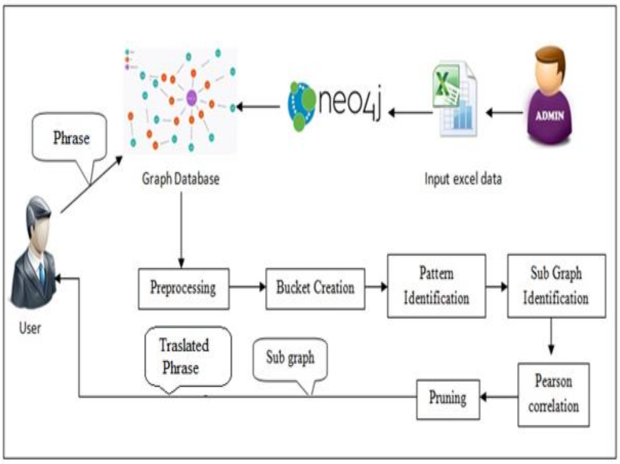

*Figure 1: System Overview*

### LAYER 1: GRAPH DATABASE GENERATION

This layer System accepts input of phrase and alterative phrase placed in excel sheet. Graph is been degenerated for each verdict indicating phrase and their alternative phrase with corresponding weight. Graph database is been generated on neo4j and entities are been represented on graph.

### LAYER 2: PHRASE PREPROCESSING

Here this layer System accepts user requirement for alternative phrases. Subgraph is been extracted from database. Following this graph is been sent for preprocessing where every phrase is been processed. Four major task are been performed segmentation, Token generation, Stop word elimination, stemming.

- ✓ Segmentation of Sentences: Finding boundary mark to separate found sentences from eliminated from phrases. Terms like “is”, ”the”, “a” are been eliminated. This process reduces time complexity of processing.
- ✓ Stemmer**:** This process a Stemmer is been applied to reduce words to their root word .like {compute, computing, computational} are been reduced to compute suffix like “ing”, “ed” are been eliminated from phrases.
- ✓ Token Generation: process to generate phrases from input words.
- ✓ Elimination of Stop words: Words having no meaning or lesser meaning are been

### LAYER 3: CREATION OF BUCKET

Similarity search is been enhanced with generation of matrix representing combination of words and phrases. This process every word from third place is been splitted to generate bucket of words for similarity search. A single vector termed as bucket is been used to hold this word. Process is illustrated as Aurangabad given to creation of bucket {Aur, Aura, Auran, Aurang, Auranga, Aurangab, Aurangaba, Aurangabad}.

### LAYER 4: PATTERN EVALUATION

Most Vital Layer in procedure where core procedure is been executed.

Pattern of Phrases are been evaluated with power set procedure applied to phrases. A power set generated is subset of all entities from singular set to union generated set.

As such Powerset of {P,Q,R}
{}
{P}
{Q}
{R}
{P, Q}
{P, R}
{Q, R}
{P, Q, R}

The above set is been used to find relation among nodes from graph database to perform subgraph matching.

### POWER SET PROCEDURE

A subset can be represented as an array of boolean values of the same dimension because the set, referred to as a characteristic vector. Each and every boolean worth suggests whether the corresponding element within the set is present or absent in the subset.

Above power set procedure for {P,Q,R} can correspond to as below mentioned.

[0, 0, 0] = {}
[1, 0, 0] = {P}
[0, 1, 0] = {Q}
[0, 0, 1] = {R}
[1, 1, 0] = {P,Q}
[1, 0, 1] = {P, R}
[0, 1, 1] = {Q, R}
[1, 1, 1] = {P,Q, R}

Procedure simply needs to produce array as above with binary counting.

### SUBGRAPH RECOGNITION

Subsequently subgraph identification is been done

Here all of the familiar words are been checked with the existing vertices of the graph information base and returns matched vector.

CORRELATION IDENTIFICATION

Here on this module resultant vector received from the last step is been feeding to the Pearson correlation to establish the correlation between the phrase phrases and the vertex feedback to extract the sub graph.

Here in this step a sub graph identification techniques is intensified based on the correlation of the edges with the aid of utilizing man or woman correlation procedure between the already recognized sub graphs.

Pearson correlation is been evaluated with equation (1)

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{10}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{10}\right)}\sqrt{\left(\sum y_i^2 - \frac{(\sum y_i)^2}{10}\right)}} \qquad ..(1)$$

Where x and y characterize the vectors of the semantic phrases in our case it's So this equation yields the correlation effect as 1. Correlation output nearer to 1 represents extra equivalent, Nearer to zero represents varied.

REDUCTION: PRUNING PROCESS

Right here in this step first unwanted vertices from the sub graph is been eliminated using horizontal and vertical pruning methods, headquartered on the aid of the graph vertices.

The whole proposed system is expressed mathematically with the below model.

Mathematical Model

1. Let S = { } be as a Machine translation
2. Identify input as P= {$p_1$,$p_2$,$p_3$ ….$p_n$ }
   Where $p_n$ = Phrase data
   S={$p_n$}
3. Identify $T_d$ as Output i.e. Translated data
   S= {$P_n$,$T_d$}
4. Identify process P= {$P_p$, $B_c$, $S_i$, $P_c$,$P_r$ }

Where,
$P_p$= Preprocessing
$B_c$=Bucket Creation
$S_i$= Subgraph Identification
Pc=Pearson Correlation
$P_{r=}$ Pruning
S = {$P_n$, $P_p$, $B_c$, $S_i$, $P_c$,$P_r$ ,$T_d$}

*The union of all subset of S Gives the final proposed system*.

## IV. RESULTS AND DISCUSSIONS

The proposed system of machine translation is developed on java based windows machine which uses Netbeans as IDE. For the experiments and performance evaluation of the system set of phrases are feed to the system to create a semantic graph and store it. To store the graph proposed system uses neo4j graph database. And developed system is put under hammer in many scenarios to prove its authenticity as mentioned in below tests.

*MEAN ABSOLUTE ERROR:* Mean Absolute Error (MAE) is one of the most using entities for scaling error rates of any outcomes, Here in our experiment MAE is used to check the error rates of the machine translation that yielded for the given String using graph approach.

MAE can be defined as shown below.

$$\text{MAE}= \sum_{i,j} | r_{i,j} – r'_{i,j} | \; / \text{N} \qquad \text{-------(1)}$$

Where,

$r_{i,j}$ ----represents the expected translation for the given string,

*r'i,j* -----represents the predicted translation for the given string.

*N* --- represents the number of predicted values.

MAE for different runs is tabulated in the below table 1.

| No of Runs | MAE |
|---|---|
| 1 | 0.7062 |
| 2 | 0.7124 |
| 3 | 0.65 |
| 4 | 0.62 |
| 5 | 0.52 |

*Table 1: MAE for different Runs*

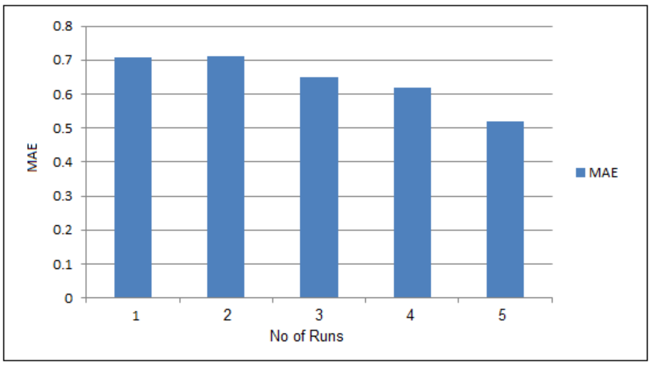

*Figure 2: Performance Graph of MAE for different query String*

The plot in the figure 2 clearly indicates that our system yields MAE of 0.64 that is good sign of machine translation in the first attempt.

## V. CONCLUSION AND FUTURE SCOPE

As the Unicode is widely using in programing languages many applications seeks the translation of languages. This is a challenging job for the application developers. Many tools are available which are translating the given phrases from one language to other. For efficient working of machine translation there is need of huge database is required for the tool to work, otherwise these tools translate the phrases non-semantically.

So proposed system introduces an idea of using NoSQL graph database to translate the string by analyzing the proper correlated patterns, which yields effective results with less size of database.

Machine translation technique can be enhance in the future by implementing the same for multiple languages in web paradigm and also this can be develop as an API.

REFERENCES

[1] WenHan Chao, ZhouJun Li, "Improved Graph-based Bilingual Corpus Selection with Sentence Pair Ranking for Statistical Machine Translation" in 23rd IEEE International Conference on Tools with Artificial Intelligence, 2011

[2] Rahul.C, Dinunath.K, Remya Ravindran, K.P.Soman, "Rule Based Reordering and Morphological Processing For English-Malayalam Statistical Machine Translation" in International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009

[3] Biswanath Barik, Sudeshna Sarkar, "Pattern based Pruning of Morphological Alternatives of Bengali Wordforms" in IEEE, 2014

[4] Fatemeh Azadi, Shahram Khadivi, "Phrase Table Pruning by Modeling the Content of Phrases" in 7th International Symposium on Telecommunications, 2014

[5] Mouiad Fadiel Alawneh, Tengku Mohd Sembok, "Rule-Based and Example-Based Machine Translation from English to Arabic" in Sixth International Conference on Bio-Inspired Computing: Theories and Applications, 2011

[6] T. B. Adji, Y. Astuti, S.S. Kusumawardani, "Statistical-based Machine Translation for Prepositional Phrase Using Link Grammar" in International Conference on Electrical Engineering and Informatics, 2011

[7] Mrs. M. Kasthuri, Dr. S. Britto Ramesh Kumar, "Rule Based Machine Translation System from English to Tamil" in World Congress on Computing and Communication Technologies, 2014

[8] Keerthi Lingam E. Rama Lakshmi L Ravi Theja,"RULE-BASED MACHINE TRANSLATION FROM ENGLISH TO TELUGU WITH EMPHASIS ON PREPOSITIONS" in IEEE, 2014