English to Telugu Rule based Machine Translation System: A Hybrid Approach

Keerthi Lingam
Assistant Professor
Department of IT
CBIT, India

E. Ramalakshmi
Assistant Professor
Department of IT
CBIT, India

Srujana Inturi
Assistant Professor
CSE Department
CBIT, India

ABSTRACT

This paper deals with adaptive rule based machine translation from English to Telugu. This approach is based on rule-based methodologies. If-then methods to select the best rules for target language in translation, Probability based appropriate word selection for a given sentence and rough sets to classify a given sentence are the approaches used in this technique. Set of production rules of English and Telugu, Training set and Dictionary for both the languages are developed for this purpose. User gives and input, which is an English sentence. The given input sentence is then tokenized into individual words. These words are tagged with their respective parts of speech. All other words that are not found in the pre-defined database are tagged using grammatical rules that are formulated. Using these POS tags, the respective word translations are retrieved from the database. These individual words are then concatenated to form a sentence that is the result of user's input.

General Terms

Artificial Intelligence, Machine Learning, Intelligent Communicating Systems.

Keywords

Machine Translation, Natural Language Processing, Rule Based Approach, English to Telugu.

1. INTRODUCTION

Language is the most extensive as well as most distinctive means of expressing ones thoughts apart from other secondary means like gestures and mime. It is also used to convey information and through which people interact. There are 7,106 spoken languages worldwide and in the era of globalization it is necessary for people to communicate in different languages. Given the amount of data available online, it is necessary to use some technique that will convert the data from a foreign language to a language that one can understand [7]. Natural language processing (NLP) is a branch of Artificial Intelligence that focuses on Machine Translation (MT). MT has become the main focus of NLP group since many years. MT deals with translating text in source language to text in target language. In the natural languages, the words in a sentence are arranged according to some predetermined rules. These rules determine if a sentence is in an acceptable form that conveys some meaning or in an unacceptable form.

Hence to build a MT system, one needs to have a clear view of the rules and grammar of the source language as well as the target language. English is a rich language and in this paper only Nouns, Verbs, Prepositions, Phrases and Infelctions are considered. This paper focused process of the MT system and the performance of it. The paper is organized as follows: Section 2 describes the existing MT systems and their merits and drawbacks. Section 3 discusses the limitations and challenges of rule based MT systems. Section 4 proposes a hybrid rule based MT system and the algorithm is discussed. Section 5 presents the implementation details and the outputs. Section 6 concludes the paper and the future work is presented.

2. EXISTING MT SYSTEMS

MT system can be developed for two specific languages and is called as a bilingual system. Multilingual system is one that is developed for more than one pair of languages. Bilingual systems are unidirectional whereas multilingual systems are designed to be bidirectional.

MT systems are classified into various categories like Rule based, example based, statistical based, hybrid based, knowledge based, principle based and online interactive based methods [9]. Rule based and statistical based methods are the earliest methods and most widely used. These approaches were used to translate the text from English to Indian languages and vice versa.

2.1 Rule Based MT Systems

Rule based MT systems were the first commercial MT systems that work on linguistic rules of source and target languages. These rules will help in arranging the translated words correctly based on the context of the sentence. Rules are applied during analysis phase, transfer phase and generation phase. This rule based system consists of various steps like syntax analysis, semantic analysis, morphological analysis, syntax generation and semantic generation. Rule based MT systems are less robust and gives good grammatical results if it finds an appropriate parse else it fails.

Rule based MT methodologies can be broadly classified as direct, transfer and Interlingua [9]. In direct methodology, there are no intermediate stages in the translation. It doesn't use any complex rules or parsing structures. This method makes use of syntactic and semantic similarities of source and target languages. Transfer methodology works in three phases namely analysis, transfer and generation. Transfer method consists of complex rules. Interlingua method works in two phases. The source text is converted into an Interlingua representation from which the text in target language is generated. In the next phase semantics of the sentence generated is analyzed.

2.2 Statistical Based MT Systems

Statistical based systems are kind of empirical MT systems which uses huge amount of information that consists of text and its translations. This approach is predicated on parallel corpora. The three key components of any statistical MT systems are language model, translation model and search algorithm.

3. LIMITATIONS AND CHALLENGES

To translate the English text to Telugu using rule based translation system understanding the structure of both the languages is important. The process of translation depends on the structure and grammar of both the languages.

3.1 Grammatical Analysis

English grammar is very rich and considerably huge in volume, hence only nouns, pronouns, verbs, articles prepositions, vibhakti (inflections) are considered in this paper. Verbs are an important part of English language and tense of a sentence can be identified by using them. Auxiliary verbs are ignored in this MT because there is no direct translation for these verbs in Telugu. A verb phrase is constructed considering the subsequent verb.

As an example consider the sentence, "Theja is going to watch a movie". It has two verbs, 'is' and 'going'. 'Is' is the auxiliary verb and 'going' is considered as the subsequent verb. Therefore 'is going' will be taken as one verb. Direct translation for 'is' in telugu is not available, so the dictionary is developed in such a manner that 'is going' is viewed as one verb phrase. Similarly in this sentence, "verbs 'to' and 'watch are also combined to one verb phrase as 'to watch. 'watch' and 'to watch' are translated differently.

'watch' is translated into telugu as 'చూడటం'(chudatam)

'to watch' is translated as 'చూడటానికి' (chudataaniki).

This 'ki' is called vibhakti in telugu. Vibhakti is one single letter or more than one letter which is added to a word in the sentence to bring out the relation with other words in the sentence. As said earlier English language doesn't have Vibhakti, so different phrases and prepositions are translated as vibhakti in telugu.

For example " Chamanthi is playing in her room" will be translated as "చామంతి తన గది లో ఆడుకుంటుంది (Chamanthi thana gadhi lo aadukuntundhi).

Here 'in' is translated as లో (lo) and added after the noun 'room' and is translated as gadhi (room) + lo (in).

While translating from one language to other, prepositions are the main issue. When there are no use of prepositions in any language, for eg. Telugu, bangle etc they will be considered as prepositional phrases and then translated using vibhakti or its equivalent in their respective languages. The dictionary that is used by translation system should be rich enough to handle them.

3.2 Structure Analysis of English and Telugu Languages

Comparative analysis of the sentence structures in English and Telugu languages is important for efficient translation. English sentences are of various types: complex sentence, compound sentence and simple sentence. Compound sentences is a combination of two or more sentences.

The language pattern for simple sentence is as follows :

Subject + Verb + Object (SVO).

For eg: Gowtham plays tennis (Gowtham + plays + tennis).

In Telugu the pattern for simple sentence is as follows :

Subject + Object + Verb.

The Telugu translation for the above sentence is as follows

గౌతమ్ టెన్నిస్ ఆడతాడు (Gowtha,+Tennis+aadathaadu)

To produce the rules for translation grammatical analysis of both the languages should be done which is similar to sentence analysis. English and Telugu languages are based on independent grammar and they need to properly mapped.

Consider the example sentence :

"Madhu reads both telugu and tamil"

Grammar pattern for this sentence is given as

n + v + (d + n' + c + n'')

where n: main noun; v: verb: d: determiner, n': noun 1; n'':noun2; c: conjunction.

Corresponding Telugu sentence would be:

"మధు తెలుగు మరియు తమిళ్ చదువుతుంధి"

Grammar pattern for this sentence is given as

$$n + (n^1 + c + n^2) + v$$

.

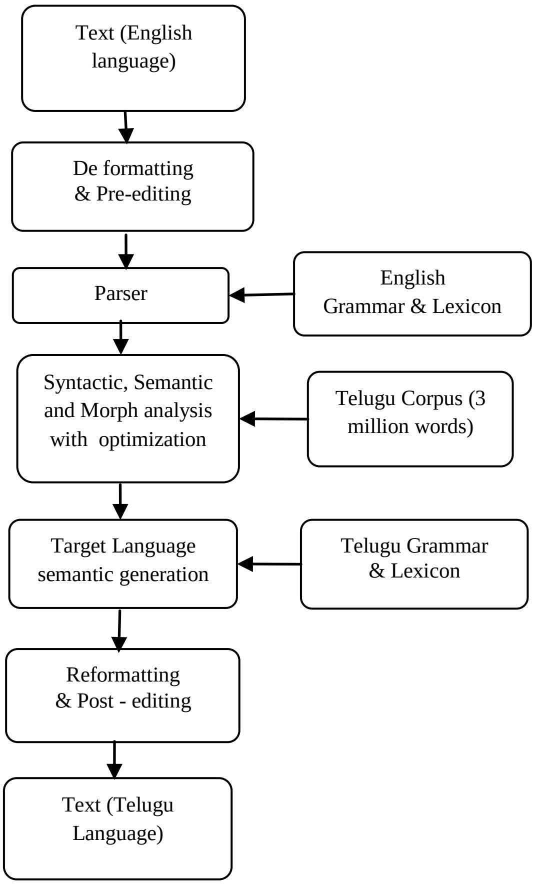

Text (English language)
De formatting & Pre-editing
Parser
English Grammar & Lexicon
Syntactic, Semantic and Morph analysis with optimization
Telugu Corpus (3 million words)
Target Language semantic generation
Telugu Grammar & Lexicon
Reformatting & Post - editing
Text (Telugu Language)

Figure 1 : Flowchart For The Proposed Technique

Morphological analysis is required to map English grammar to Telugu grammar. Hence it is important and also necessary to compare the structure of grammar of both the languages in order to achieve efficient language translation. English and Telugu are two diverse languages, that prepositions and Auxiliary verbs from English are not found in Telugu grammar. Likewise Vibhakti which are a part of Telugu grammar are not used in English grammar. So auxiliary verbs will be considered as verb phrases by taking in the subsequent verb and adding to this verb. Prepositions are considered as prepositional phrase where Vibhakti is introduced as suffix to the noun in the Telugu Sentence.

4. A HYBRID RULE BASED MACHINE TRANSLATION SYSTEM

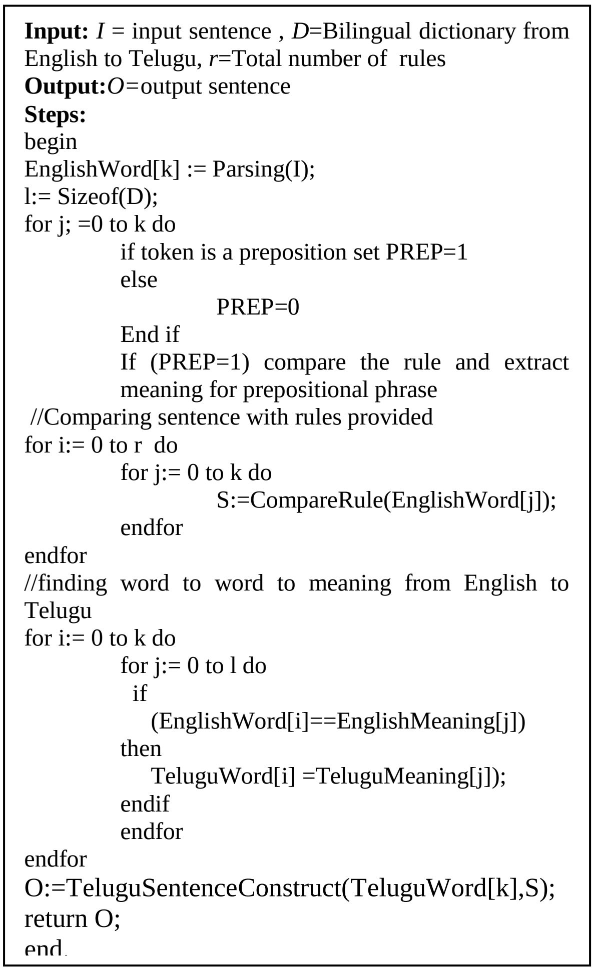
Efficient rules are framed for translating the text from English (source language) to Telugu (Target language) using hybrid rule based machine translating system. In the implementation phase a set of English sentences are translated to Telugu and these will be considered as the training set. Then after another set of unseen sentences will be given to the system to check the efficiency of translation

Initially a sentence given by the user is given as input to “explode()” function. It breaks the given string into words or tokens. Each of these words is considered individually and their corresponding tenses and words are identified. In general a sentence in English is of the form “S V O”. Rules are hence framed to convert that format into “S O V”. This process is described precisely as below.

Speaking in general terms, rule based MT generates the target text given a source text following the steps shown in Figure 1 and the algorithm is shown in Figure 2.

4.1 Algorithm for Rule Based Machine Translation

The text that is to be translated will be identified from figures and flowcharts in the Deformatting phase. The figures and flowcharts need not be translated. At the end, soon after the translation the text will be reformatted with figures and flowcharts in the Reformatting phase.



```
Input: I = input sentence , D=Bilingual dictionary from English to Telugu, r=Total number of  rules
Output:O=output sentence
Steps:
begin
EnglishWord[k] := Parsing(I);
l:= Sizeof(D);
for j; =0 to k do
        if token is a preposition set PREP=1
        else
                PREP=0
        End if
        If (PREP=1) compare the rule and extract
        meaning for prepositional phrase
 //Comparing sentence with rules provided
for i:= 0 to r  do
        for j:= 0 to k do
                S:=CompareRule(EnglishWord[j]);
        endfor
endfor
//finding word to word to meaning from English to Telugu
for i:= 0 to k do
        for j:= 0 to l do
          if
            (EnglishWord[i]==EnglishMeaning[j])
        then
            TeluguWord[i] =TeluguMeaning[j]);
        endif
        endfor
endfor
O:=TeluguSentenceConstruct(TeluguWord[k],S);
return O;
end.
```

Fig 2: Algorithm For The Proposed Technique

Vaious symbols and punctuation marks in the text need not be translated and those will be taken care in the pre editing and post editing phases of translation. As previously discussed, syntactic and semantic and morphological analysis are done on the text in source language. Based on the output of these phases and based on the rule identified, corresponding rule for the target language that is Telugu will be picked out.. In the following stage, grammatical representation of the Telugu sentence will be generated. The next phase is where the exact words in target language based on the context and meaning of the input sentence is to be identified. Contextual semantic and syntactic generation will reduce the ambiguity.

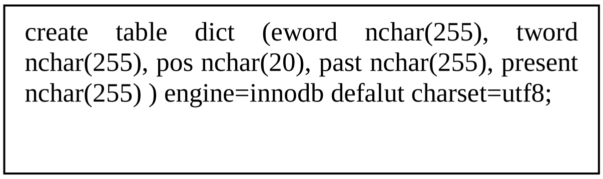
4.2 Production Rules

The set of production rules designed for translating the text from English to telugu is shown in Table 1. As the training set increases these rules can also be increased. This work is limited only to work on Nouns, pronouns, verbs, prepositions, articles and adjectives. Dictionary is also expandable. Anytime new words along with their information is is added into it. Words are then stored with their attributes.

5. IMPLEMENTATION

This MT system converts a given English sentence into its corresponding Telugu sentence using HTML(Front-end), PHP(Middleware), MySQL(Back-end) technologies. All the words are identified irrespective of whether they are listed in the dictionary or not. An input is accepted from user, which is an English sentence. The given input sentence is then tokenized into individual words. These words are tagged with their respective parts of speech. All other words that are not found in the database are tagged using grammatical rules that we formulated. Using these POS tags, their respective word translations are retrieved from the database. These individual words are then concatenated to form a sentence which is the result of user’s input.

Database that supports UTF-8 encoding is created through the following code given in Figure 3.

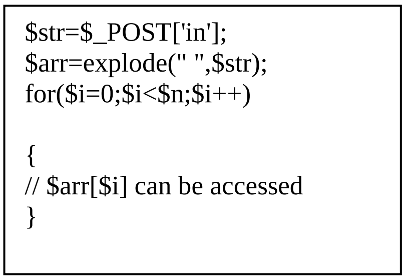


```
create table dict (eword nchar(255), tword nchar(255), pos nchar(20), past nchar(255), present nchar(255) ) engine=innodb defalut charset=utf8;
```

Figure 3: Sample code for Database Creation

5.1 Tokenization

As shown in Figure 4, $str is a PHP variable that takes as input through $_POST['in'] function, where ‘in’ is the input given by the user. Explode() function splits the user string into an array named $arr. The for loop is the used to access each of the array elements.



```
$str=$_POST['in'];
$arr=explode(" ",$str);
for($i=0;$i<$n;$i++)

{
// $arr[$i] can be accessed
}
```

Figure 4: Sample Code for Tokenization

Table 1. Production Rules for English to Telugu Translation

PRODUCTION RULES			
ENGLISH PATTERN		**TELUGU PATTERN**	
PR$_1$	s → n + v + n^1	PR’1	*s* --> *n* + *n*^{1} + *v*
	Rama killed Ravana		రామ + రావణని + చంపాడు
PR$_2$	p + v	PR’$_2$	p + v
	we + were dancing		మేము + నృత్యం చేస్తున్నాము
PR$_3$	n + v	PR’$_3$	n + v
	The gold + glitters		బంగారం +మెరుస్తుంది
PR$_4$	p + d + v	PR’$_4$	p + d + v
	we + all + eat		మనం +అందరం +తింటాము
PR$_5$	d + art + n	PR’$_5$	d + art + n
	that + is a + dog		అది+ ఒక+ కుక్క
PR$_6$	n + v + (p + n^1)	PR’$_6$	n + (p + n^1) + v
	John + took + (our + photo)		జాన్ +(మా +ఫోటో)+ తీసాడు
PR$_7$	n + v + (p + art + adj + n^1)	PR’$_7$	n + (p + art + adj + n^1) + v
	teacher + told + (us + an + interesting + topic)		టీచర్ +(మాకు+ ఒక+ ఆసక్తికరమైన+ విషయం)+ చెప్పారు
PR$_8$	p + v + (n + d + n^1)	PR’$_8$	p + (n + d + n^1) + v
	we + visited + (Tajmahal + last + year)		మేము +(తాజమహల్+ పోయిన+ సంవత్సరం)+వెళ్ళాము
PR$_9$	p + v + adv	PR’$_9$	p + adv + v
	she + was writing + then		ఆమె+ అప్పుడు+ రాసింది
PR$_{10}$	v + p + d + adj + n	PR’$_{10}$	p + d + adj + n + v
	give + them + some + light + work		వాళ్లకి+ కొంచెం+ తేలిక+ పని+ ఇవ్వండి
PR$_{11}$	p + v + prep + d + n	PR’$_{11}$	p + n + prep + v
	He + is + in + the + park		అతడు + తోట + లో + ఉన్నాడు
PR$_{12}$	p + v+ v + prep + n	PR’$_{12}$	p + n+ prep + v
	I + am + walking + in + park		నేను + తోట + లో + నడుస్తున్నాను
PR$_{13}$	n + v + prep + d + n^1	PR’$_{13}$	n + n^1 + prep + v
	Book + is + under + the + table		పుస్తకము + బల్ల + కింద + ఉంది
PR$_{14}$	p + v + adj	PR’$_{14}$	p + adj
	It + is + good		అది + మంచిది
PR$_{15}$	p + v + v^1 + prep + n	PR’$_{15}$	p + n + prep + v
	I + am + walking + beside + river		నేను + నది + పక్కన + నడుస్తున్నాను

5.2 POS Tagging

As the words are stored in a database along with their parts of speech, gender and tense, extraction of tense is done as follows:

```
$result=mysql_query("select pos from dict where eword='$trim’",$con);

$row=mysql_fetch_array($result);

array_push($b,$row['pos']);
```

The above is a mysql code embedded in PHP. dict is the table in database that stores all the words and the afore mentioned fields. The parts of speech of the input word is extracted and input to a new array, $b. This array stores the parts of speech of the entire sentence as array elements. array_push() is a function used to insert elements into this array.

5.3 Rules Framing

It is based on the parts of speech that rules are written for getting a proper Telugu sentence as output. Some of the basic rules for translation are shown in Figure 5.

```
$c=implode(" ",$b);
if(strcmp($c,"pro v v")===0)
{
array_push($out,$t[0]);
array_push($out,$t[2]);
array_push($out,$t[1]);
}
if(strcmp($c,"pro v det n")===0)
{
        array_push($out,$t[0]);
        array_push($out,$t[2]);
        array_push($out,$t[3]);
        array_push($out,$t[1]);
}
```

Figure 5: Sample Rules

In the above code, $c is a string of parts of speech obtained using implode () function. This function unites the array elements into a string ($c in this case). Strcmp () function compares the obtained parts of speech and re-arranges the order of parts of speech along with the corresponding Telugu words.

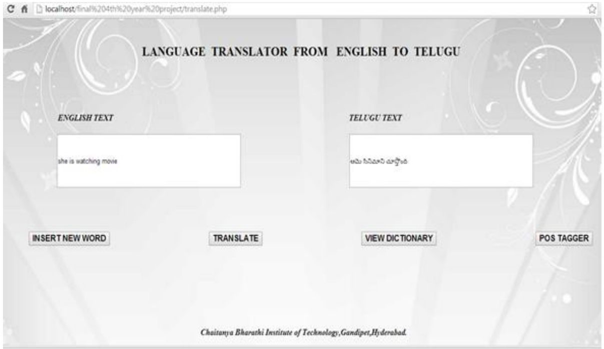
Apart from the above mentioned steps, the major step of displaying the text is done using UTF-8 (Universal Transformation Code) encoding. The words are displayed in Telugu by setting the content type of HTML as shown in Figure 6.

```
header('Content-type: text/html; charset=UTF-8') ;
```

Figure 6: Sample Code to Display Telugu Text in HTML

5.4 Results

The first phase in machine translation system is giving text input, which is shown in Fig 7. The parts of speech tagging is shown in Fig 8. This system also allows the user to enter the new words into the dictionary. Users can have a brief glance at the dictionary to obtain meanings of simple words instead of sentences as shown in Figure 9.



localhost/final%204th%20year%20project/translate.php

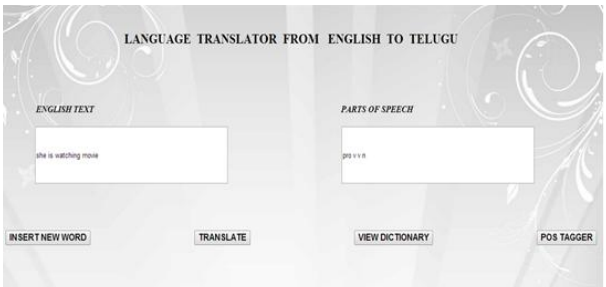
LANGUAGE TRANSLATOR FROM ENGLISH TO TELUGU

ENGLISH TEXT: she is watching movie

TELUGU TEXT: ఆమె సినిమాని చూస్తోంది

INSERT NEW WORD
TRANSLATE
VIEW DICTIONARY
POS TAGGER
Chaitanya Bharathi Institute of Technology,Gandipet,Hyderabad.

Figure 7: Translation Process



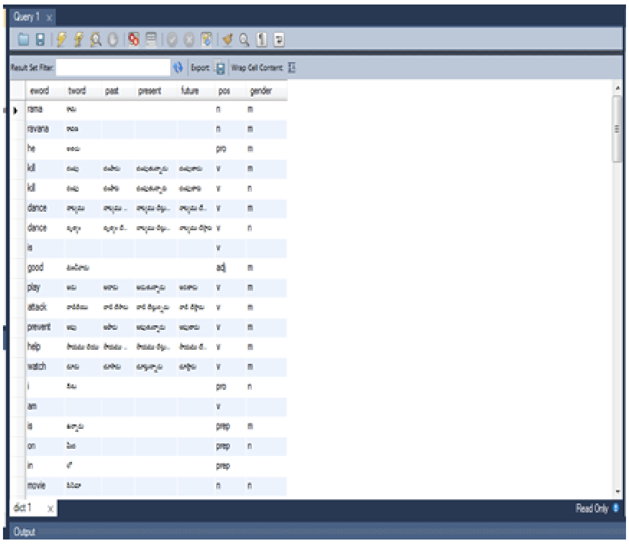
LANGUAGE TRANSLATOR FROM ENGLISH TO TELUGU

ENGLISH TEXT: she is watching movie

PARTS OF SPEECH: pro v v n

Figure 8: POS Tagging

The internal database created in WAMP server can be displayed using MySQL workbench, as it supports displaying UTF-8 encoded data, unlike WAMP server.



eword	tword	past	present	future	pos	gender
rama	[illegible]				n	m
ravana	[illegible]				n	m
he	[illegible]				pro	m
kill	[illegible]	[illegible]	[illegible]	[illegible]	v	m
kill	[illegible]	[illegible]	[illegible]	[illegible]	v	n
dance	[illegible]	[illegible]	[illegible]	[illegible]	v	m
dance	[illegible]	[illegible]	[illegible]	[illegible]	v	n
is					v	
good	[illegible]				adj	m
play	[illegible]	[illegible]	[illegible]	[illegible]	v	m
attack	[illegible]	[illegible]	[illegible]	[illegible]	v	m
prevent	[illegible]	[illegible]	[illegible]	[illegible]	v	m
help	[illegible]	[illegible]	[illegible]	[illegible]	v	m
watch	[illegible]	[illegible]	[illegible]	[illegible]	v	m
i	[illegible]				pro	n
am					v	
is	[illegible]				prep	m
on	[illegible]				prep	n
in	[illegible]				prep	
movie	[illegible]				n	n

Figure 9: Dictionary Database

This system is tested on two data sets, each for training and testing the efficiency. 92% efficient translation is achieved and other sentences were partially correct.

6. CONCLUSION

A new approach based on rule based MT for translating text from English to Telugu is proposed and implemented in this work. This work can be improved further by adding extra rules to the set of the rules that are available now. The Transliterate part can be enhanced by using efficient bilingual dictionary and the search methods. Even the morphology of the translated language has to be improved by taking a variety of example sets that they represent the defined rules. People who take up this project in future can also develop the performance based report to improve the performance and efficiency.

7. REFERENCES

[1] Akshar Bharati, Vineet Chaitanya, Amba P. Kulkarni and Rajeev Sangal, " ANUSAARAKA: Machine Translation in Stages", A Quaterly in Artificial Intelligence, Vol. 10, No. 3, July 1997

[2] Sanjay Kumar Dwivedi and Pramod Premdas Sukhdev, "Machine Translation in Indian Perspective", Journal of Computer Science, june 2010.

[3] Latha R Nair and David Peter S, "MAchien Translation system for Indian Languages", IJCA, Vol 39, No. 12012

[4] Sugata Sanyal and Rajdeep Borgohain, "Machine Translation Systems in India", arXiv, April 2010.

[5] Judith Francisca and Md Mamun Mia, "Adapting Rule Based Machine Translation From English To Bangla", IJCSE, Vol 2. No. 3 Jun-Jul 2011.

[6] Sitender and Seema Bawa, "Survey of Indian Machine Translation Systems," IJCST, Vol 3, Issue 1, Jan-Mar 2012

[7] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

[8] Lewis, M. Paul, Gary F. Simons, and Charles D.Fennig(eds.). 2014. Ethnolouge: Languages of the world, seventeenth edition, Dallas Texas: Sil International. http://www.ethnologue.com.

[9] Antony, P. J. "Machine Translation Approaches and Survey for Indian Languages." *Computational Linguistics and Chinese Language Processing Vol* 18 (2013): 47-78.