A Hybrid Rule-Based Machine Translation System from English to Telugu

Keerthi Lingam, Ramalakshmi E\textsuperscript{1}, and L. Ravitheja\textsuperscript{2}

\textsuperscript{1} Chaitanya Bharathi Institute of technology/Dept of IT, Hyderabad, India Email: \{lkeerthi, eramalakshmi\}@cbit.ac.in
\textsuperscript{2} Gudlavalleru Engineering College/Dept of IT, Gudlavalleru, India Email: lravitheja93@gmail.com

Abstract—This paper deals with adaptive rule based machine translation from English to Telugu. This is a proposed approach and it is based on rule-based methodologies. Various approaches like If-then based methods to select the appropriate rule for target language, Rough sets to classify a given sentence into a particular class and probability to decide whether the word selected for Telugu is appropriate for the given sentence or not are used in this method. Set of production rules, training set for English sentences and English to telugu dictionary are developed for this purpose.

Index Terms—Artificial Intelligence, Machine translation, Natural Language Processing, Rule based approach.

I. INTRODUCTION

Machine Translation, a sub-discipline of Artificial Intelligence is the process of translating text from one human language(source language) into another human language (Target language). This is considered as a difficult problem because processing human languages requires work at several levels and many complexities arise at each level. It is complex but it is not impossible because the number of words in any natural language is finite, very large but finite. India is a multilingual country where there are more than 30 languages and around 2000 dialects used for communication, and there is a big requirement for Interlingua translation that helps people in communicating easily and share their ideas. This paper deals with work that translates text from English to Telugu. To proceed with the Machine translation from English to Telugu, one needs to understand the sentence structure and grammar of both the languages. Telugu is rich in morphology and is an agglutinative language than English. Only a part of the English grammar like Nouns, verbs, phrase, prepositions and inflections are considered in this work. The focus of the work is on the process of machine translation and the effectiveness. The paper is organized as follows: Section II discusses the Existing work and classification of machine translation. Section III explains the problem statements. Proposed algorithm and solution is analyzed in Section IV, implementation details and expected results are provided in Section V. Finally, Section VI concludes the work and provides scope for future works.

II. EXISTING TRANSLATION SYSTEMS AND CHALLENGES IN MACHINE TRANSLATION

Telugu is one of the most spoken Dravidian languages almost the number summing up to 76 million. Telugu script is written same as English, left to right on a page from top to bottom. Any sentence is partially ended with a single bar(purna viramam) and permanently ended with double bars (Deergha viramam). The pattern consists of 16 vowels and 41 consonants and three vowel modifiers. Telugu has three genders, Masculine, Feminine and neutral. The grammar rules of Telugu are obtained from Paninian’s concepts. The translation of Telugu language to English is done based on SOV-Subject Object verb.
Anusaaraka is the translation system developed by IIT Kanpur and IIIT Hyderabad in 1995 and works on language knowledge, word knowledge and statistical knowledge [1]. The output given by this system can be edited if there are any grammatical errors. The MANTRA is another tool to translate text from English to Hindi and this approach is formalized on Tree Adjoining Grammar (TAG). [2]. Anubharathi-II is built using Example based and corpus based approach. [3].

A. Classification of Machine Translation Systems

Machine translation systems are broadly classified into three categories: Direct machine translation, Rule based machine translation and Empirical based machine translation. Rule based machine translation can be further classified into Transfer based approach and Interlingua based approach. [4]. Empirical based machine translation can be further classified into Statistical machine translation and Example based machine translation. Apart from these approaches, few hybrid machine translation systems also showed fruitful results.

B. Rule Based Machine Translation System

This approach depends on built in production rules and dictionaries that are created for each language pair. Machine translation system parses the text that has to be translated and a transitional representation is created from which the target language text can be generated. This whole process goes through various phases like syntactic, semantic and morphological analysis that uses large set of rules. Rule based systems produces good grammatical results if it finds a parse else it fails. They are less robust. The error rates, which are now 3-4%, can also be improved and there is lot of cope for improvement. [5] The direct language translation totally depends on how well the source-target language dictionaries are built and morphological analysis. Whereas in Interlingua based translation source language text will be translated into intermediatly form called interlingua. This system can be used to translate the text in one source language to many target languages. Complexity of the system is the main drawback.

C. Direct Machine Translation and Empirical Based Machine Translation

Direct Machine translation is based on translating words directly from source language to target language without having to convert into intermediate representation. Empirical approach uses large amount of raw data that consists of text and its translations. This approach is based on parallel corpora.

III. PROBLEM STATEMENT

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

A. Grammatical Analysis

English grammar is very huge in volume, so let us only consider few essential parts, verbs, prepositions, vibhakti (inflections), phrases. Verbs are very important in English and one can identify the tense of a sentence using them. In this translation, auxiliary verbs are ignored because there is no direct translation to this verb in Telugu. A verb phrase is constructed considering the subsequent verb. For example the English sentence, “John is going to play chess” has two verbs, ‘is’ is the auxiliary verb and ‘playing’ is the subsequent verb. This will be considered as one verb like ‘is playing’. There is no direct translation for ‘is’ in telugu, so the dictionary is also developed in the same way and ‘is going’ is considered as one verb phrase. Similarly verbs ‘to’ and ‘school’ are also combined to one verb phrase as ‘to play’. ‘Play’ and ‘to play’ are translated differently. Play is translated into telugu as ‘ఆడటం’ (aadatam) and to play is translated as ‘ఆడట”ఆడటం’ (aadatanaiki). 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 “ Mary is studying in her room” will be translated as “మయిల్ కు రంచిందు చదువుకంం” (Mary thana房间lo chadvukuntundhi). Here ‘in’ is translated as ‘lo’ (lo) and added after the noun ‘room’ as ‘roomlo’ (roomlo). 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.

B. Structure analysis of English and Telugu languages

Comparative analysis of the sentence structures in English and Telugu languages is important for efficient translation. Sentence is of three types: Simple, complex and compound sentence. Compound sentences can be a combination of two or more than two sentences. The language pattern for simple sentence is as follows Subject + Verb + Object (SVO). For eg: She plays cricket (she + plays + cricket). In Telugu the pattern for simple sentence is as follows Subject + Object + Verb. The Telugu translation for the above sentence is as follows

అంటే వీరు ఆడంతూ (aame+cricket+aadutundhi)

Like sentence analysis, grammatical analysis should also be done for both languages to produce rules for translation. English and Telugu languages have their own set of grammar and we need to properly map them. Consider the following English sentence, “I play both tennis and chess”. Grammar pattern for this sentence is as follows

\[ n + v + (d + n^1 + c + n^2) \]

n: main noun, v: Verb, d: Determiner, n\(^1\): Noun\(^1\), c: Conjunction, n\(^2\): Noun \(^2\)

Grammar for corresponding telugu sentence “I play both tennis and chess” (nenu tennis mariyu chess, rendu aadathanu) is given as follows:

\[ n + (d + n^1 + c + n^2) + v \]

Morphological analysis should be done to map English to Telugu grammar. So it is necessary to compare the grammar structures of both the languages for efficient language translation. There is much dissimilarity between the two languages, as the auxiliary verb and prepositions of English are not found in Telugu grammar. Similarly, vibhakti are not used in English grammar. So we considered auxiliary verb as verb phrase by adding this verb to the subsequent verb in the sentence, Prepositions are considered as prepositional phrase or vibhakti that is added to the noun as suffix in the Telugu sentence.

IV. PROPOSED SYSTEM

In this work, rules are designed for efficient machine translation from English to Telugu. During the implementation phase, a set of English sentences will be translated into Telugu and they can be considered as training set. After that another set of new sentences can be given to the system for translation to check the efficiency.

A. Algorithm For Rule Based Machine Translation

The algorithm for this machine translation is shown in Fig 1. and the flow chart for whole process is shown in Fig 2. Different phases represent different steps of the translation process. DE formatting and reformatting is the phase in machine translation that will identify the text from figures, flowcharts etc. that does not require any translation. Once the translation is done reformatting should be done to get the exact output along with all figures and flowcharts. Pre-editing is to fix up the punctuation marks, symbols etc which do not require any translation whereas post editing is to ensure that the translation is done efficiently. As discussed earlier morphological, syntactic and semantic analysis will be done on the source text. Based on the output of this phase and the rule identified, corresponding rule for the target language that is Telugu will be identified. In the next stage, internal representation of the sentence in Telugu will be generated. The next phase, contextual semantic and syntactic generation will help in finding the exact words in Telugu and frame the sentence.

B. Production Rules

The set of rules used for translating the training set is shown in Table 1. These set of rules can extended as the training set increases. Production rules for English and its corresponding Telugu sentences are shown in the table. All these rules are predefined. Dictionary used in this translation is expandable. The words in the dictionary are classified into different classes and information about each word is also stored. Words are
stired in the dictionary with their attributes. More number of words will be included in the dictionary as more no of sentences are added to the training set.

**Fig 1: Algorithm For The Proposed Technique**

```
Input: I = input sentence, D=Bilingual dictionary from English to Telugu, r=Total number of rules
Output: O=output sentence

Steps:
begin
InputWord[0] := Parsing(I);
D := Sizeof(D);
foreach sentence with rules provided do
for i:= 0 to r do
   for j:= 0 to k do
      S := CompareRule(EnglishWord[i],
   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[i];
      endif
   endfor
endfor
O := TeluguSentenceConstruct(TeluguWord[k],S);
return O;
end
```

**Fig 2: Flowchart For The Proposed Technique**
TABLE 1: PRODUCTION RULES FOR ENGLISH TO TELUGU TRANSLATION

<table>
<thead>
<tr>
<th>PRODUCTION RULES</th>
<th>ENGLISH PATTERN</th>
<th>TELUGU PATTERN</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR₁</td>
<td>s → n + v + n'</td>
<td>PR'₁</td>
</tr>
<tr>
<td></td>
<td>I am playing + chess</td>
<td>ఇది + వన + వన్న కమ్మాను</td>
</tr>
<tr>
<td>PR₂</td>
<td>p + v</td>
<td>PR'₂</td>
</tr>
<tr>
<td></td>
<td>we + were dancing</td>
<td>వెడ + వెడ + కుంటి సంపనన మన</td>
</tr>
<tr>
<td>PR₃</td>
<td>n + v</td>
<td>PR'₃</td>
</tr>
<tr>
<td></td>
<td>The gold + glitters</td>
<td>గాలం + యుర + కుంటి సంపనన మన</td>
</tr>
<tr>
<td>PR₄</td>
<td>p + d + v</td>
<td>PR'₄</td>
</tr>
<tr>
<td></td>
<td>we + all + eat</td>
<td>మనం + అందరం + ఆంటం మన</td>
</tr>
<tr>
<td>PR₅</td>
<td>d + art + n</td>
<td>PR'₅</td>
</tr>
<tr>
<td></td>
<td>that + is a + dog</td>
<td>అక్కడ ఉంటా + వన + వన్న కమ్మాను</td>
</tr>
<tr>
<td>PR₆</td>
<td>n + v + (p + n')</td>
<td>PR'₆</td>
</tr>
<tr>
<td></td>
<td>John + look + (our + photo)</td>
<td>జంపా + లాంక + (మీ + ఫోటా) మన</td>
</tr>
<tr>
<td>PR₇</td>
<td>n + v + (p + art + adj + n')</td>
<td>PR'₇</td>
</tr>
<tr>
<td></td>
<td>teacher + told + (us + an + interesting + topic)</td>
<td>టిచ్చారు + తొలి + (మీ + అన + విశేషాత్మక + ప్రాంకమ్) + మన</td>
</tr>
<tr>
<td>PR₈</td>
<td>p + v + (n + d + n')</td>
<td>PR'₈</td>
</tr>
<tr>
<td></td>
<td>we + visited + (Tajmahal + last + year)</td>
<td>వెడ + విసిట్ + తాజమహాల + లాసి + సంవసలం + మన</td>
</tr>
<tr>
<td>PR₉</td>
<td>p + v + adv</td>
<td>PR'₉</td>
</tr>
<tr>
<td></td>
<td>she + was writing + then</td>
<td>సై + వాటిని + తొనడ + మన</td>
</tr>
<tr>
<td>PR₁₀</td>
<td>v + p + d + adj + n</td>
<td>PR'₁₀</td>
</tr>
<tr>
<td></td>
<td>give + them + some + light + work</td>
<td>గివ్ + టిండో + సాం + లిగ్ + ము అవసరం</td>
</tr>
</tbody>
</table>

V. IMPLEMENTATION AND RESULTS

This translation technique can be implemented by using any programming language like C, C++, Java, VB etc. The few phases that we implemented are written in Java. Words are tokenized and is done by split() method in Java. Next phase is to define the phrases of the sentence and this step helps in finding out the relative Telugu word from the dictionary. When an appropriate word is found in the dictionary, the information associated with it is accessed to find if it is a correct word for the Telugu sentence. Probability of the occurrence of that word in the given sentence is used to select the word. PoS tagging that is done in the beginning will help construct the sentence in Telugu with proper tenses.

This system can be tested using two data sets. One data set for training the system and other data set for testing. Approximately 92% result is expected for 100% true translation. Other sentences could be partially correct. These results are expected in comparison with the existing techniques.

VI. CONCLUSION

This paper deals with a new approach for machine translation from English to Telugu and this is based on Rule based machine translation. The work that is done in English to Telugu translation is very little and among the different approaches that are available, this is unique. We are working on writing effective set of rules and building a dictionary that is rich in words, so that our translation system can deal with any kind of sentences. Classification of words is another challenging task, which we are working on. We want to extend this work to identify the phrases and idioms and translate them in a better manner instead of direct translation.

REFERENCES


