Conversion of Text from English to Indic Language (Telugu) using Binary Tree Traversal

Devaki Pendlimarri
Dept. of Computer Science and Engineering
Viswam Engineering College
Madanapalle, Andhra Pradesh, India
mrsdevaki02@gmail.com

Paul Bharath Bhushan Petlu
Dept. of Computer Science and Engineering
Viswam Engineering College
Madanapalle, Andhra Pradesh, India
trishipaul@gmail.com

Abstract— **Machine Translation (MT) is aimed to enable a computer to transfer the Natural Language utterances either in text or speech from one language to another preserving its meaning and interpretation. There are several paradigms from the beginning of the MT, including word-to-word direct translation, rule-based transfer approach, inter-lingua approach and knowledge-based machine translation (KBMT). Many MT systems on the market use a translation scheme called a sentence structure conversion scheme. This paper presents an approach which is a combination of word-to-word direct and rule-based transfer approaches to translate the English sentence into the Indic languages and from one Indic language to another Indic language.**

Keywords-machine translation; natural language processing; telugu; english; text translation; conversion of text; binary tree traversal; knowledge based machine translation;

I. Background

With the recent advances in the Information Technology, the Machine Translation Systems (MTS) aimed to enable a computer system to transfer the text or speech from one natural language into another natural language preserving its meaning and interpretation. There are several paradigms from the beginning of the Machine Translation (MT) including word-to-word direct translation, rule-based transfer approach, inter-lingua approach and knowledge-based machine translation. Many MTS on the market use a translation scheme called a sentence structure conversion scheme. INDIA is a country where the people of different region speak different language. Though the people speak different languages, Hindi is the national language of INDIA. Equally with the national language, English is also treated as an official language in INIDA. In INDIA, the Department of Information Technology initiated the TDIL (Technology Development for Indian Languages) with the objective of developing Information Processing Tools and Techniques to facilitate human-machine interaction without language barrier; creating and accessing multilingual knowledge resources; and integrating them to develop innovative user's products and services. This paper presents a tree traversal approach to translate the text in English to Indic languages by using some set of rules.

II. Introduction

INDIA is a country which is like a mini world having the people with different language, culture, tradition, customs etc. across 500km distance. The languages spoken by the Indians are called as the Indic languages. In the state of Andhra Pradesh, people speak the language called "Telugu", Tamilnadu state is "Tamil" etc. like that each state will have a different language as its official language. At the same time all the states are having the language "Hindi" as another official language as it is the National language of INIDA. English is also another official language in the country. When the people move to different place over 500km there language is becoming a barrier for communication. All the Indic languages are having almost the similar structure and semantics. The differences between the Indic languages are minor which can be noticed and solved.

Generally, the translation process in traditional MT has three modules, known as, analysis, transfer and generation. The analysis module deals with the transformation of the source language utterances into a predefined format of internal representation, through morphological processing, parts-of-speech (POS) tagging, syntactic parsing, semantic analysis, etc. The transfer module works to convert the representation of the source language into that of the target language. The generation module is concerned with the derivation of target utterances from the representation, observing necessary syntactic, semantic and pragmatic constraints.

In this paper, we are presenting a method of translating an input text in English into the Indic languages without changing its meaning and interpretation.

III. Methodology

In this paper, we are presenting a methodology to translate the input text which is in English into the Indic languages (can be converted into other languages too depending upon the syntactic and semantic rules of the language). This methodology consists of two steps: *preprocessing* and the *translation* as shown in Figure 1. The MTS pro- cess starts with the preprocessing, which in turn consists of two steps: *Text splitter* and *Tagger.*

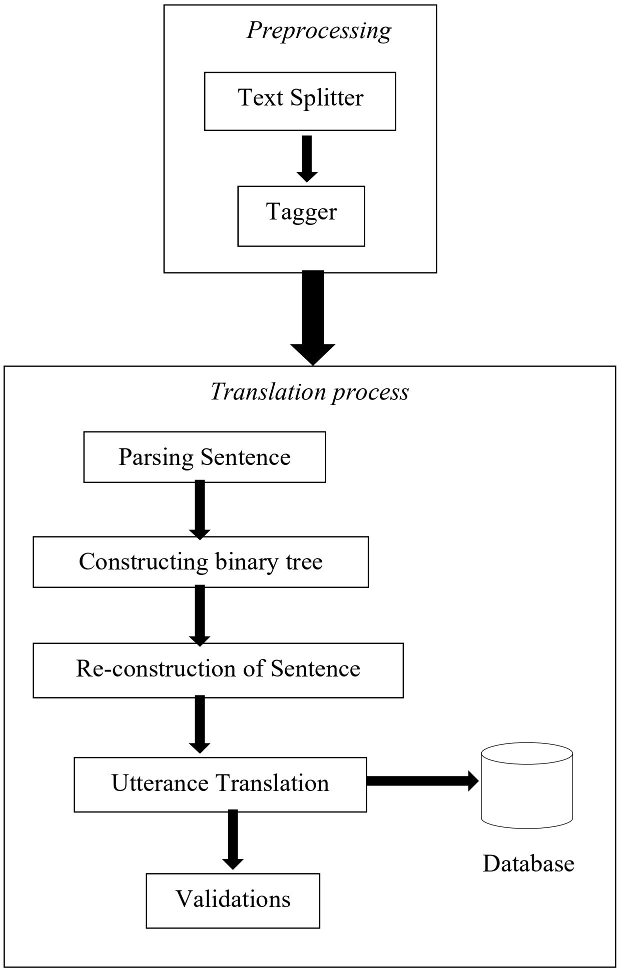


Figure 1. Process of MTS

The *text splitter* will split the given input text into number of sentences using the delimiter "period or dot". The *Tagger* will tag e*ach* sentence with the Tagger. The translation process is applied to each tagger to translate each sentence. The *translation* process consists of several steps, such as, parsing, constructing a binary tree, reconstruction of a sentence into the target language, utterance translation into the target language and validations.

IV. Implementation

We are implementing the above methodology using JAVA as frontend and ORACLE as its backend. The steps are given below along with some examples. The input source language is taken as English and the output target language is taken as Telugu. The output target language may be taken as any Indic language. This can also be applied to other languages too with the notice of the syntactic and semantic rules of that language.

A. Parsing

The text in English is a set of sentences which are delimited by a period. Each sentence is a set of words delimited by a space. Parsing is the first step where the given input text is divided into number of sentences. The MTS is applied on each sentence to translate it to the target language. In this process the first step is to parse the sentence to get the utterances based on the delimiter "space".

B. Constructing a binary tree

After parsing, the next step is constructing a binary tree which is considered as the most important step in translation process. In this step, each word is inserted into the binary tree depending upon the type of word (Parts-of-Speech) with the help of predefined set of rules and preference value of each word. While constructing the binary tree, we also note the vital information presented in Table 1. This vital information is essential because words in the Indic languages are changing basing on the vital information.

Table 1. Vital Information

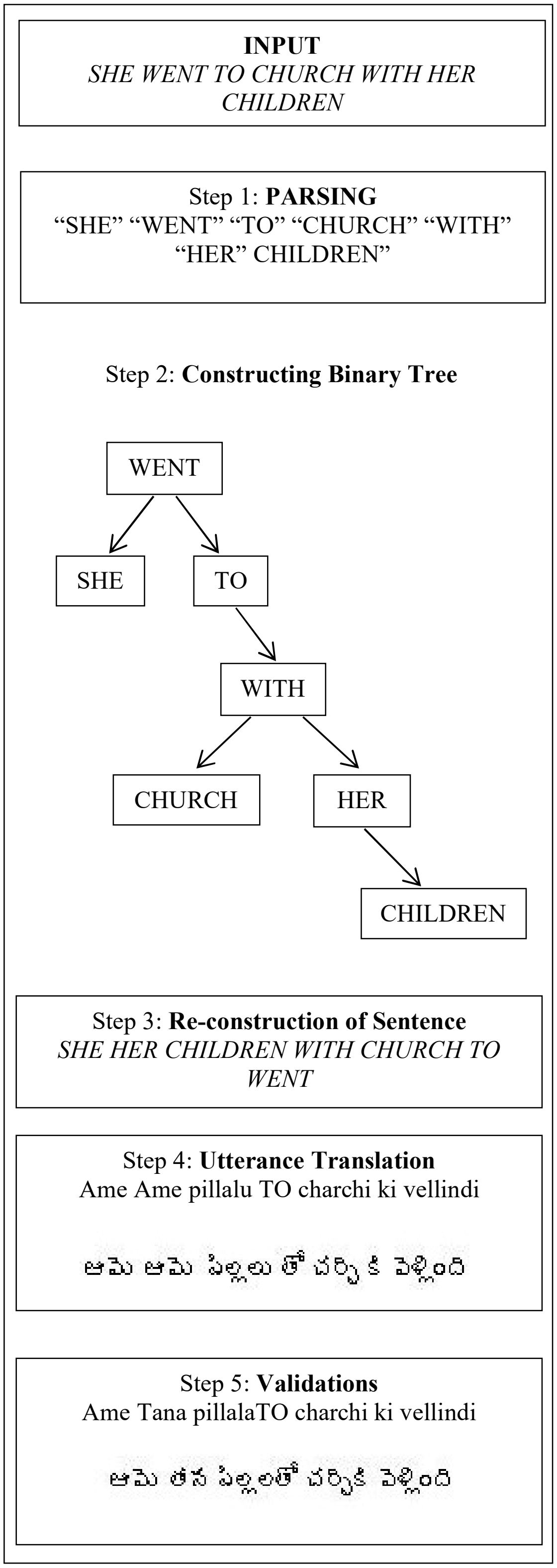
S. No.	Vital Information
1	Singular and Plural
2	Masculine and Feminine
3	Countable and Uncountable

C. Reconstruction of sentence

From the binary tree constructed in the previous step, we reconstruct the sentence into the language we have chosen. This process uses the tree traversal techniques and the set of predefined rules. The utterances are still in the source language. The sentence reconstructed is with the English words but in the form of target language based on the syntactic and semantic rules of the target language.

D. Utterance translation

After reconstructing the sentence into the syntactic and semantic rules of the target language, the next step is to translate the actual meaning of each utterance into the target language. The meaning of each utterance is stored in the database. A mapping of each utterance of the source language sentence is made into the target language sentence with the help of *vital information* in step 2.



INPUT
SHE WENT TO CHURCH WITH HER CHILDREN

Step 1: **PARSING**
"SHE" "WENT" "TO" "CHURCH" "WITH" "HER" CHILDREN"

Step 2: **Constructing Binary Tree**

Step 3: **Re-construction of Sentence**
SHE HER CHILDREN WITH CHURCH TO WENT

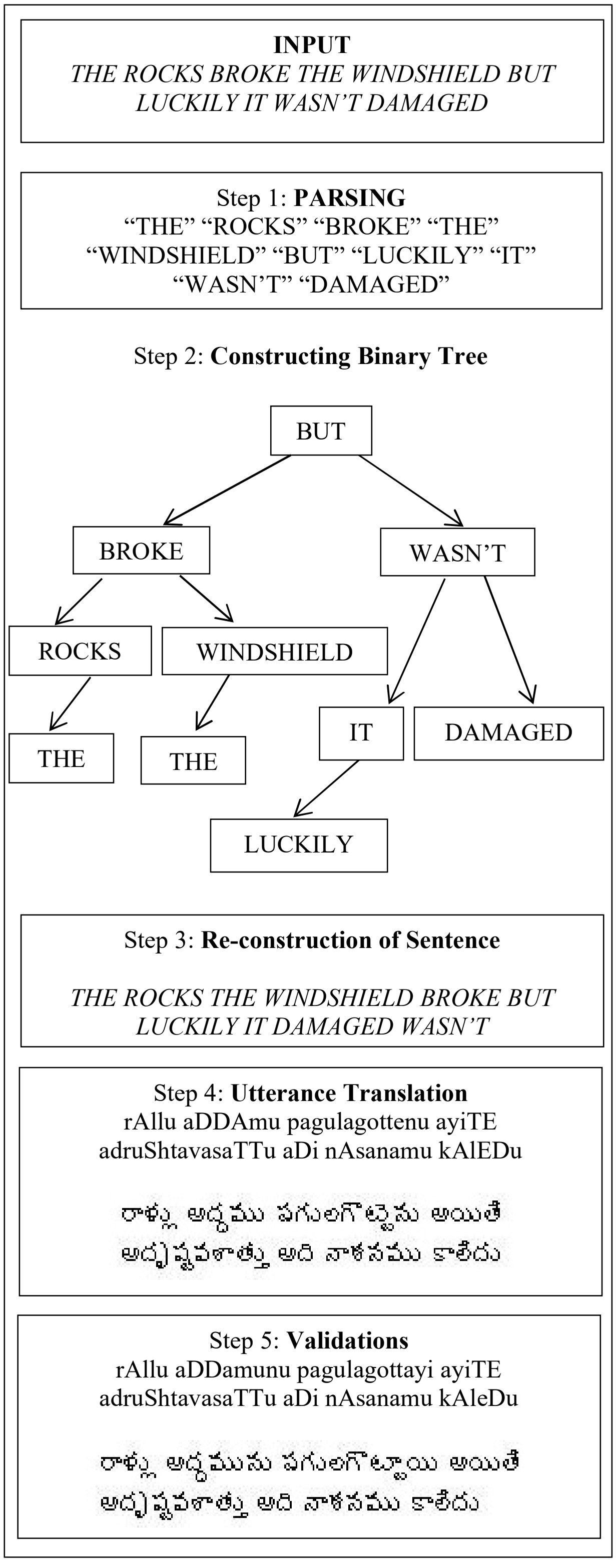
Step 4: **Utterance Translation**
Ame Ame pillalu TO charchi ki vellindi

ఆమె ఆమె పిల్లలు తో చర్చి కి వెళ్లింది

Step 5: **Validations**
Ame Tana pillalaTO charchi ki vellindi

ఆమె తన పిల్లలతో చర్చికి వెళ్లింది

Figure 2. MTS process of example 1



INPUT
THE ROCKS BROKE THE WINDSHIELD BUT LUCKILY IT WASN'T DAMAGED

Step 1: **PARSING**
"THE" "ROCKS" "BROKE" "THE" "WINDSHIELD" "BUT" "LUCKILY" "IT" "WASN'T" "DAMAGED"

Step 2: **Constructing Binary Tree**

Step 3: **Re-construction of Sentence**

THE ROCKS THE WINDSHIELD BROKE BUT LUCKILY IT DAMAGED WASN'T

Step 4: **Utterance Translation**
rAllu aDDAmu pagulagottenu ayiTE
adruShtavasaTTu aDi nAsanamu kAlEDu

రాళ్ళు అద్దము పగులగొట్టెను అయితే
అదృష్టవశాత్తు అది నాశనము కాలేదు

Step 5: **Validations**
rAllu aDDamunu pagulagottayi ayiTE
adruShtavasaTTu aDi nAsanamu kAleDu

రాళ్ళు అద్దమును పగులగొట్టాయి అయితే
అదృష్టవశాత్తు అది నాశనము కాలేదు

Figure 3. MTS process of example 2.

E. Validations

This is the most important step to get the out with the same meaning and interpretation of the source language sentence into the target language sentence. We will have a predefined set of rules based on the target language. Every will have its own set of rules and regulations which is known as *grammer.* These predefined set of rules are constructed based on these rules. These rules are applied to get the right interpretation in the target language.

V. RESULTS AND ANALYSIS

The algorithms of the methodology are verified and validated with some example and got the good result. The given input text in English was translated into the target language Telugu without changing its meaning and interpretation. Some of the examples are given below.

Example 1: The MTS process of this example is shown in Figure 2. The input text in English was taken as *"SHE WENT TO CHURCH WITH HER CHILDREN"*. By parsing this text is broken into number of words as "SHE" "WENT" "TO" "CHURCH" "WITH" "HER" "CHILDREN". In the second step these words are inserted into a binary tree one by one based on the predefined set of rules. While inserting the words into the binary tree some vital information which is required in further steps is noted. In the third step, the words in the binary tree are reconstructed into a sentence into the target language based on the predefined set of rules. Here, we have taken the Indic language "Telugu" as the target language. This target language can be any language provided that the respective predefined set of rules is defined. In the fourth step, the MTS will convert each word into the target language by connecting to the database. These converted words will not make a right sentence in the target language. In the fifth step, by applying the validation rules of the target language, the converted words are joined to form a right sentence without changing the meaning and interpretation of the source language sentence into the target language.

Example 2: The MTS process of this example is shown in Figure 3. The input text in English was taken as *"THE ROCKS BROKE THE WINDSHIELD BUT LUCKILY IT WASN'T DAMAGED"*. By parsing this text is broken into number of words as "THE" "ROCKS" "BROKE" "THE" "WINDSHIELD" "BUT" "LUCKILY" "IT" "WASN'T" "DAMAGED". In the second step these words are inserted into a binary tree one by one based on the predefined set of rules. While inserting the words into the binary tree some vital information which is required in further steps is noted. In the third step, the words in the binary tree are reconstructed into a sentence into the target language based on the predefined set of rules. Here, we have taken the Indic language "Telugu" as the target language. This target language can be any language provided that the respective predefined set of rules is defined. In the fourth step, the MTS will convert each word into the target language by connecting to the database. These converted words will not make a right sentence in the target language. In the fifth step, by applying the validation rules of the target language, the converted words are joined to form a right sentence without changing the meaning and interpretation of the source language sentence into the target language.

VI. CONCLUSION AND FUTURE WORK

In this paper we have presented a methodology to translate the given input text in English into the target language without changing its meaning and interpretation. This translation is mainly based upon the construction of binary tree from the input sentence and reconstruction of sentence into target language from binary tree. Presently we are implementing the system in JAVA for the Indic languages only. This also can be implemented for other languages too. In future, we would like to implement this MTS to translate a sentence given in any language to any other language.

REFERENCES

[1] Abbot, R. J. 1983, *"Program Design by informal English descriptions"*, Communications of the ACM, vol. 26, 882-894

[2] Arthern, P. J. (1978). *Machine Translation and computerized terminology systems: a translator's viewpoint. Translating and the computer: Proceedings of a Seminar* (pp. 77-108). London.

[3] BLIS (1998). *Bilingual laws information system* (BLIS). Info. Tech. and Resources Unit. Admin. Division, Dept. of Justice, HK Government. Information available from http://www.justice.gov.hk/

[4] E. Minkov, K. Toutanova and H. Suzuki, *"Generating complex morphology for machine translation"*, In ACL 07: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 128-135, Prague, Czech Republic. Association for Computational Linguistics, 2007

[5] G S Ananda Mala and Dr. G V Uma, *Object Oriented Visualization of Natural Language Requirement Specification and NFR Preference Elicitation*, IJCSNS International Journal of Computer Science and Network Security, Vol.6, No.8, August 2006

[6] Indian Language Technology Proliferation and Deployment Centre ILTP-DC available at *URL:* http://www.tdil-dc.in/

[7] K. Papineni, S. Roukos, T. Ward and W-J. Zhu, *"BLEU: A Method for Automatic Evaluation of Machine Translation"*, in Proceedings of ACL 2002, Philadelphia, PA., pp 311-318, 2002

[8] Nirenburg, S., Beale, S., & Domashney, C. (1994). *A full-text experiment in example-based machine translation.* International Conference on New Methods in Language Processing (pp 78-87). Manchester

[9] NIST machine translation scoring and tests is available at *URL:* http://www.nist.gov/speech/tests/mt

[10] P. Koehn and H. Hoang *"Factored Translation Models"*, In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868-876, Prague, June 2007

[11] S. Goldwater and D. McClosky. *"Improving statistical MT through morphological analysis"*, In Proceedings of Human Language Technology Conference and Conference on empirical methods in Natural Language Processing, pages 676-683, Vancouver, British Columbia, 2005

[12] Sumita, E., Lida, H., & Kohyama, H. (1990). *Translating with examples: A new approach to machine translation*. TMI'90 (pp. 203-212). Texas

[13] Yarowsky, D. (2000). Word-sense disambiguation. In R. Dale, H. Moisl and H. Somers (Eds.), *Handbook of Natural Language Processing*, 629-654. New York: Marcel Dekker

[14] Y. S. Lee, K. Papineni, S. Roukos, O. Emam and H. Hassan. *"Language model based Arabic word segmentation".* In E. Hinrichs and D. Roth, editors, Proceedings of the 41st Annual Meeting of the Association for Computational Linguistic, 2003

[15] http://en.wikipedia.org/wiki/Georgetown-IBM_experiment

[16] Google translation: https://translate.google.com/?hl=en&tab=TT