Oriental Journal of Computer Science & Technology Vol. **3**(2), 317-322 (2010)

Rule based approach for english to Sanskrit machine translation and synthesizer system

D.T. MANE[1], P.R. DEVALE[1] and S.D. SURYAWANSHI[2]

[1]Bharati Vidyapeeth's College of Engineering, Pune (India).
[2]Rajarshi Shahu College of Engineering, Pune (India).

(Received: May 22, 2010; Accepted: June 25, 2010)

ABSTRACT

The area of Artificial Intelligence is very useful in providing people with a machine, which understands diverse languages spoken by the common man. It presents the user with an interface, with which he feels more comfortable. Since, there are many different languages spoken in this world, we are constantly in need for translators to enable people speaking different languages to share ideas and communicate with one another. English is the global language .The most of the information is available in English. The India is a country which has several regional languages. Sanskrit is the mother of all native language of India. A great storage of knowledge with subjects like medicine, mathematics, Geography, Geology, Astronomy, philosophy and many others is kept alive and fresh Sanskrit lore for thousands of years.English to sanskrit translator and Synthesizer is very useful to people in India , sentence in English is translated in to sanskrit using rule based approach and from sanskrit it is easier to transform in to native languages.

Key words: Rule based dictionary approach, Parser, bilingual dictionary, Formant Synthesizer.

INTRODUCTION

The module present concerns with the Machine Translation domain of Natural Language Processing. This area of Artificial Intelligence is very useful in providing people with a machine, which understands diverse languages spoken by common people. It presents the user of a computer system with an interface, with which he feels more comfortable. Since, there are many different languages spoken in this world, we are constantly in need for translators to enable people speaking different languages to share ideas and communicate with one another.

The India is a highly multilingual country with eighteen constitutionally recognized languages and several hundred dialects & other living languages. Even though English is understood by very less people in India. It continues to be the de-facto link language for administration & education system. Internet is media for information retrieval &information is available in English on internet, so there is need of translator which can convert English sentences in to native languages. Sanskrit act as a interlingua for translation to and from Indian languages.

The Sanskrit language is basically the language of the ancient India and considered as the mother language from which all other Indian languages evolved. In this time Sanskrit language is a dead language. But it is recognized in the Indian constitution of 1950 because Sanskrit is related and associated with the religion and literature of India.

Sanskrit language is used in some other states of India as Haryana, Delhi, Rajasthan, Jammu and Kashmir and the out of India in worldwide countries as UK, USA, Bangladesh Canada, UAE, Singapore, Kenya, Fiji and Malaysia. Sanskrit language belongs to the Indo-Iranian family of languages. It's an Indian historical language.

For translation we decode the meaning of the source input text in its entirety, the translator must interpret and analyse the text, a process that requires deep knowledge of the grammar, semantics, syntax, idioms, etc., of the source language and Target Language. The translator needs the same in-depth knowledge to re-encode the meaning in the target language.

For translation generally rule based, Statistical based and Example Based approach used. and Rule Based methods for machine translation is divided in to another three types

1. Transfer based machine translation: - type of machine translation based on idea of interlingua and is currently one of the most widely used area. It is necessary to have an intermediate representation that captures the "meaning" of the original sentence in order to generate the correct translation.
2. Dictionary based machine translation: - Machine translation can use dictionary based approach , which means that the words will be translated as they are by a dictionary.
3. Interlingual based machine translation :Machine translation can use dictionary based approach mean the text to be translated, is transformed into an interlingual, i.e. source or target-language-independent representation. The target text e is then generated out of the interlingua

In this paper we consider dictionary rule based approach is for translation and syhthesizer . In dictionary based approach words are stoed in Database dictionary and when we got input then English sentence are separated from sentence by tokenization then morphological analysis is done, after getting the words its search into English dictory and according to word its category e.g. (noun, verb) is assigned.

If we compare the Grammar for both English and Sanskrit then English sentances always in order of subject-verb-object format while sanskrit has free word order. For e.g.the order of english sentence (ES) and its equivalent translation in sanskrit sentence (SS) is given below.

ES: He read book(SVO)
SS : Saha pustkam Pathati.(SOV)
OR
Pustkam Saha Pathati.(OSV)
OR
Pathati Pustkam Saha

Thus sanskrit sentence can be written using SVO, SOV and VOS order.

Speech is the primary means of communication between people. Speech synthesis, automatic generation of speech waveforms .Synthesis procedure consists of two main phases. The first one is text analysis, where the input text is transcribed into a phonetic or some other linguistic representation, and the second one is the generation of speech waveforms, where the acoustic output is produced from this phonetic and prosodic information. These two phases are usually called as high- and low-level synthesis.

System design

Design is a process through which requirement are translated into a representation of software. The careful analysis of this module and its possible solutions leads to the following design of the system. The sections to follow give the total design as well as the details of the design of individual modules. The details are illustrated using diagrams and examples.

The main idea behind dictionary based Machine Translation is that input text sentence can be converted in to output sentence by carrying out the simplest possible parse ,replacing source word with their target language equivalents as specified in a bilingual dictionary, and then using grammer rules of target langauge re-arranging their order .

The overall system is diagrammatically shown below

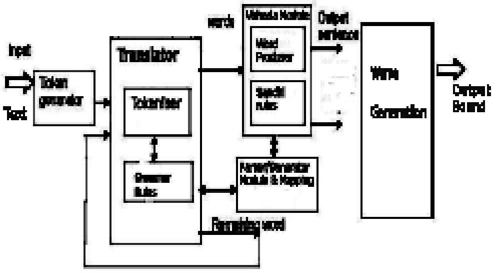


Fig. 1: Basic Block Diagram of Translator & Synthesizer

Following Steps are involve in translation and Synthesizer

- Token Generator: This module splits the given sentence into chunks of strings delimited by spaces. These strings may be simple words or compound words coalesced by the rule of English Grammar. By applying the rules of English grammar assign appropriate category to words like (noun, verb, noun phrase etc.) Generate a parse tree using grammar rules of source language.
- Vichcheda Module: The vichcheda module gets help from the set of transducers to identify words and forms words through the word generator. The word generator in turn takes the help of the sandhi rules module wherever necessary. The remaining string after the basic word is generated is sent back to the vichcheda module.
- Translator: This module performs the actual translation. The input to this module is the parse. It also interacts with the parser/ generator module to get the parse of each word. It then generates appropriate equivalents in English for the morphological details of each word and ultimately presents the sentence in the correct order.
- Find translation of all English words in to Sanskrit dictionary Rearrange the words in Sanskrit using its grammar rule to format and a meaningful statement. Parser Generator Module contains a set of transducers built for individual Sanskrit words and transforms strings to partial words, which are used by the vichcheda module and dictionary based approch. It also gives the parse of the words, which are used by the sentence former to give the output in a structurally correct sentence.
- After getting the target text it converts into the wave and it plays using some output devices i.e speakers. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity

Consider the example

A man goto school

Narah pathashalam gachati

Step1: Token generatorThis module separete each word from sentence according to english grammer.it also cosider the spac e between two words.

A.	Man	go to	school
Token 1	Token2	Token3	Token4

Step2: Parse or syntax tree represent the syntatic structure of a sentence according to general grammer. In a tree, the interior nodes are labeled by non-terminals of the grammar, while the leaf nodes are labeled by terminals of the grammar. A program that produces such type of ordered tree is called a parser.

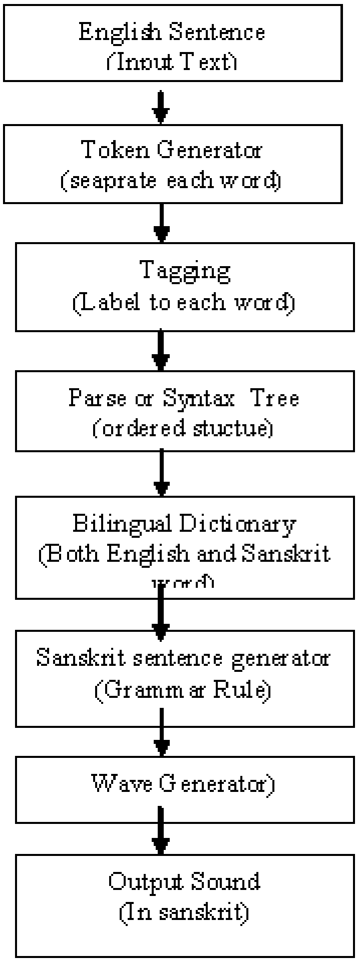


Fig. 2: System for machine translation

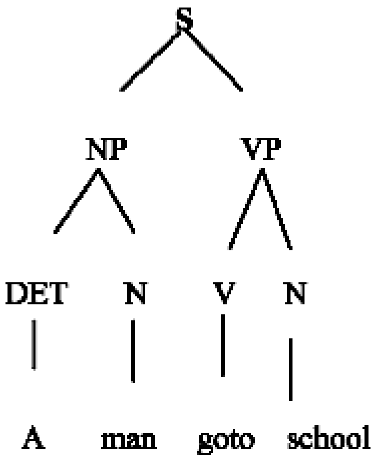


Fig. 3: Parse tree of source sentence

Create a parse tree of source statement using parser

Step3: After creating parse tree we get the each word with proper tagging then find the meaning of each word in english dictionaryand Sanskrit dictionary.If meaning of word is not available ic dictionary its gives error massage.

Step4:Considering grammer rules of sanskrit language generate the sanskrit sentence and after getting the prper output sentence it forward to next synthesizer module..

Destination sentence:- Narah shaakam khadati

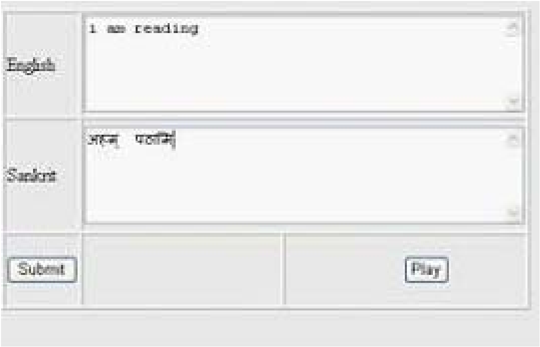
A man	:- Narah	: Subject.
To School	:- Pathashalam	: Object
go	:- Gachati	:Verb

Step 5: After geeting the sanskrit sentence apply some techniques and it converted into wave form and then this waves given to out put devices like as speakers

EXPRIMENTAL

The objectives described above may be implemented using different appraches. We use here only one dictionary base approach. The separate lexicons for English sentence and Sanskrit sentence may be maintained in a database with morphological details stored in the form of logics in the programming language used. A bilingual dictionary will also have to be maintained in this case.

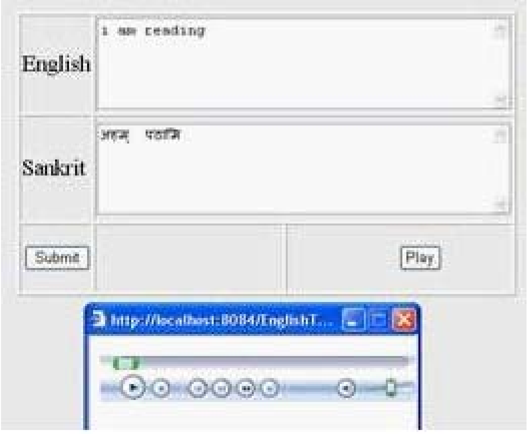
Rules are formed for tokenization and parsing ,Tokenization and parsing is implemented using java languageof any other logical programming language.i.e Prolog.


English: i am reading

Sankrit: अहम् पठामि

Fig. 4: Expected output screen after translator Module

Exepected output screen after getting translation means convert the english sentence to sanskrit sentence is


English: i am reading

Sankrit: अहम् पठामि

Fig. 5: Expected output screen after nthesizer Module

Exepected output screen after synthesizer means convert the sanskrit sentenceinto waveform and it play using oupt devices i.e speakers.

Dictionary based approach

In rule base dictionary approach for translation limitation or weak points are

1. It requires much more linguistic knowledge.
2. It is impossible to write rules that cover all a language.
3. In case of modifying some rules, It does not only change the incorrect sentence in to correct sentence, It may be further affecting on the correct sentence also.
4. Common error occur in machine translation is one is choosing incorrect meaning and another one is incorrect sequence.
5. The translation process involves the many different rules interacting in many different ways. Due to which it is hard to extend or modify.
6. This approach is designed with translation in one direction ,between one pair of languages in mind it is not conductive to the development of genuiely multilingual.
7. Text normalization challenges, Most text-to-speech systems do not generate semantic representations of their input texts, Deciding how to convert numbers is another problem
8. Text-to-phoneme challenges: Speech synthesis systems use two basic approaches to determine the pronunciation of a word based on its spelling, a process which is often called text-to-phoneme or grapheme-to-phoneme conversion (phoneme is the term used by linguists to describe distinctive sounds in a language).

CONCLUSION

The conception of the idea of building the present paper had purely research motives. Though translation of sentences from English to Sanskrit is the main aim of the present paper, performing tokeniser,sandhi vichcheda and performing the morphological parsing simultaneously receives a greater part of the attention in building the system. The aim was to implement the idea presented in my project is implementing the translator using some basic dictionary approach.

Dictionary based approach of rule base method of MT is possible it required dictionaries of both the languages along with morphological databases of both the languages.

And Speech synthesis has been developed steadily over the last decades and it has been incorporated into several new applications. For most applications, the intelligibility and comprehensibility of synthetic speech have reached the acceptable level.Formant TTS voices are typically not as natural-sounding as concentrative TTS voices and I use formant approach synthesizer for this stage , but both provide capabilities that prerecorded audio cannot; most notably, the ability to present unbounded, dynamic information to the user.

This tool will very useful for the sharing the worldwide knowledge with Indian.

REFERENCES

1. R.M.K. Sinha and A. Jain, AnglaHindi: An English to Hindi Machine-Aided Translation System. In “Proceedings of MT SUMMIT IX”, New Orleans, Louisiana, USA (2003).
2. Aparna Subrmanian “ Sanskrit to English Translator”(Dissertation submitted at master in computer science at DAV,Indore).
3. Daniel Juratsky ,James martin “ speech & language processing –an introduction to natural language processing (2000).
4. Vimal Mishara, R.B. Mishara “Study of Example Based English to Sanskrit Machine Translation” department of computer engineering ,Institute of technology,Banaras Hindu University, Varanasi India.
5. D.J. Arnold, “Machine translation an introductory guide”.
6. M. R. Kale, “A Higher Sanskrit Grammar”, 4th Ed, Motilal Banarasidas Publishers Pvt. Ltd., (2005).
7. Macdonnel, “A Sanskrit Grammar for Students”, 3rd Ed, Motilal Banarasidas Publishers Pvt. Ltd, Books (2003)
8. Machine Translation: An Introductory Guide.

By Doug Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys and Louisa Sadler (1994).

9. An overview of machine translation, by John Hutchins (University of East Anglia, United Kingdom: updated January 2005),

10. “Speech and Language Processing- an introduction to Natural Language Processing” by Daniel Jurafsky and James Martin, reprint (2000).

11. Natural Language Processing by Aksar Bharati, Vineet Chaitanya, Rajeev Sangal.

12. Introduction to Computer Theory” -second edition by Daniel I.A.Cohen

13. A Higher Sanskrit Grammar by M. R. Kale

14. Sandhi Viveka” by A. Varadaraj

15. Text to Speech Synthesizer” by Paul Taylor.