Machine Translation from English to Hindi

G. Suryakala Eswari[#1], N.V.S.Sowjanya[*2], P.Surya Prabhakar Rao[#3]

[#]*Information Technology, JNTUK, Kakinada.*
[1]*padmakala1@gmail.com*
[2]*sowjanya828@gmail.com*
[3]*suryaprabhakar.p@gmail.com*

Abstract— *Statistics show that only ten percent of Indian population is aware of English language to the extent that they can speak and understand spoken English. Since a majority of the population do not have working knowledge of English and since English is world communication medium it becomes imperative to understand the current happenings in the world. A way to achieve this to be able to translate English language into Hindi which may be used to bridge the gap. Many research organizations in India and abroad have started developing translating systems for translating English into various Indian languages. Machine translation, as the name says, is the process of translating using a machine, such as a computer.*

The proposed system is direct translation system. It is based on the bilingual dictionaries. In the proposed system, the source language sentence is given as input. The processing involves splitting sentence into words and looking up in the dictionary for translation. If the word is not available in the dictionary, then transliteration is done.

Keywords— **Machine Translation, Transliteration, Morphological, statistical machine translation, communication.**

I. Introduction

Machine translation can be defined as translation from one language to another language. And also it is a subfield of computational linguistics it uses software to translate text or speech from one language to another language. It is the part of natural language processing.

A. Machine Translation (MT) in India

MT is important technology for localization and particularly relevant in linguistically diverse like India. Human translation in India is rich and important and it works on different fields like arts, science, and philosophy. It finds so many applications mainly in the administration, media, education and business.

It helps people to understand language very easily. The language that is to be translated is called source language, after translation the obtained language is called target language.

B. Why MT is popular in India?

India has linguistically rich area. It has twenty two constitutional languages which are written in ten different scripts. Hindi is the official language of the union. However, English is very widely used in the media, commerce, science, technology, and education. India’s national language is Hindi. Sixty to seventy percent of the population understands Hindi and five percent of the population speaks English. So, in India there is big market for translation from English to Hindi.

As is clear, the market is the largest for translation from English to Indian languages, primarily Hindi. Hence English to Hindi translation systems have great demand. In English to Hindi machine translation system, English is source language and Hindi is target language. Most of the MT systems use rule based translation and statistical machine translation techniques.

C. Difference between English and Hindi

English is a highly positional language and default sentence structure is subject, verb, object (SVO).In SVO languages, the verb comes between the subject and the object.

Hindi language has rich morphology in that, it has free word order and default structure is subject, object and verb (SOV). In SOV languages the verb comes at the end of the basic clauses.

SVO languages generally have prepositions, whereas SOV languages have postpositions.

D. Why MT is hard?

There are several structural and stylistic differences among the languages, which makes translation a difficult task. In translation, it is very difficult to handle the words, which have multiple meanings. So, more intelligent selection methods need to be designed for MT.

E. Architecture

The machine translation system has three architectures, namely, (i) Direct translation, (ii) Transfer, and (iii) Interlingua as depicted in Figure 2.1

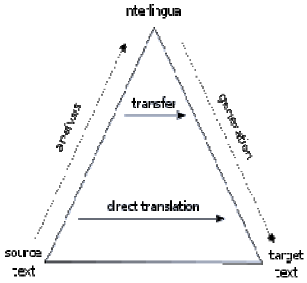

Fig 2.1 Architectures for machine translation

F. Direct translation

In direct translation, we proceed word by word through the source language text, translating each word as we go. We make use of no intermediate structures, except for shallow morphological analysis. Each source word is directly mapped onto some target word. Direct translation is based on large bilingual dictionary. Each entry in the dictionary can be viewed as a small program, whose job is to translate one word at a time. After the word is translated, simple reordering rules can apply.

G. Transfer

The structure of input shall be to make it conform to the rules of the target language. This can be done by applying contrastive knowledge, that is, knowledge about the differences between two languages. The transfer model parses the source language and generates a phase to actually create the output sentence. Thus on this model, MT involves analysis and generation, where transfer bridges the gap between the output of the source language parser and the input to the target language generator. MT systems can also use some times combination of the direct and transfer approaches, using large bilingual dictionaries with taggers and parsers.

H. Interlingua

Interlingua as a process of extracting the meaning of the input and then expressing that meaning in the target language. We can create this Interlingua representation from source language text by using the semantic analyzer.

I. Statistical Machine Translation

Statistical machine translation can be defined as generation of translations using statistical methods based on bilingual text, Corpa. In statistical machine translation, we observe the importance of both faithfulness and fluency. Statistical MT is used to build probabilistic models of faithfulness and fluency and then combine these models to choose the most probable translation. The faithfulness and fluency are quality metrics that can be used to model the translation from a source language sentence S to target sentence T.

II. EXISTING SYSTEMS IN INDIA

A survey of the machine translation systems that have been developed in India for translation from English to Indian languages and among Indian languages is made. And the MT software is used in field testing or is available as web translation service. These systems are also used for teaching machine translation to the students and researchers. Most of these systems are in the English-Hindi or Indian languages-Indian Languages (for example Hindi to Tamil). The translation domains are mostly used to translated government documents/reports and news stories. There are a number of other MT systems developed in India. Many of these Systems cover other Indian languages besides Hindi.

India has a big list of spoken languages. At least, 30 different languages have been identified. In India the two mostly used languages are Hindi and English for official communication for the national government. Additionally, it uses 22 scheduled languages which are languages that can be officially adopted by different states for administrative purposes, and also as a medium of communication between the national and the state governments, as also for examinations conducted for national government service.

A. Anglabharathi and Anubharathi

The Anglabharathi project was launched by Professor R. M. K.Sinha at the Indian Institute of Technology, Kanpur in 1991 for machine aided translation from English to Indian languages. The system has been applied in several domains such as public health campaign, routine office-correspondence, technical-manual development etc.

The first prototype was built for English to Tamil in 1991 and later built for English to Hindi translation. Anglabharati is a pattern of directed rule based system with context free grammar like structure for English (source language). It generates a ‘pseudo-target’ (Pseudo-Interlingua) applicable to a group of Indian languages (target languages) such as Indo-Aryan family (Hindi, Bengali, Assamese, Punjabi, Marathi, Oriya, and Gujarati etc.), Dravidian family (Tamil, Telugu, and Kannada& Malayalam) and others.

Anglabharathi-II uses a generalized example-base (GEB) for hybridization besides a Raw Example-Base (REB). During the development phase, it was found that the modification in rule-base is difficult and may result in unpredictable results.

The English to Hindi version is named AnglaHindi, based on the Anglabharathi machine aided system.

B. Anusaraka

Anusaraka can be defined as a language help between Indian languages, using principles of Paninian grammar. Anusaraka is a machine aided translation system among Indian languages and it has been built with funding from the TDIL project. Anusaraka is a Language access rather than a machine translation system in true sense. Anusaraka have been built to translate from Telugu, Kannada, Bengali, Marathi, and Punjabi to Hindi. Alpha versions of all of these have been released so that their field testing can be carried out. The beta-versions are expected to be released soon. It is domain free but the system has mainly been applied for translating children’s stories. The focus in Anusaraka is not mainly on machine translation, but on language access between Indian languages using principles of Paninian Grammar (PG). Exploiting the close similarity of Indian languages, an Anusaraka essentially maps local word groups between the source and target languages. Where there are differences between the languages, the system introduces extra notation to preserve the information of the source language. Thus, the user needs some training to understand the output of the system.

III. PROPOSED SYSTEM

The proposed system is direct translation system. It is based on the bilingual dictionaries. In the proposed system, the source language sentence is given as input. The processing involves splitting sentences into words and arranging words in sentences, based on grammar rules (SVO is arranged as SOV; the sentences starting with words “what”, “where”, “why”, “how”, and”when” are arranged in appropriate places in the sentence. If the word is present in dictionary, translation is done. If the word is not available in the dictionary, then transliteration is done.

The Indian language scripts are syllabic and consist of V, CV, and CCV, CCCV type of units where C is a constant and V is a vowel. To render the syllables on the computer screen, Unicode and Unicode rendering engine available in Window xp and Linux operating systems are used. Since English is a positional language and the default sentence structure is **S**ubject **V**erb and **O**bject (SVO) and Hindi is a morphological language with free word order and the default structure is **S**ubject **O**bject **V**erb (SOV). Additional steps convert the order of sentence in Hindi after the translation and transliteration.

The implement phase consists of coding the modules in java with eclipse as Integrated Development Environment (IDE). Test cases are written and tested successfully.

A. Transliteration

Transliteration is one of the phases of Machine Translation. Transliteration is defined as the task of transcribing a word or text from one writing system into another writing system. Cognates (the words derived from another language) and Named Entities (NE) such as the person names, names of places, organizations are the types of words that need to transcribe into another writing system. Transliteration editors are essential for keying-in Indian language scripts into the computer using QWERTY keyboard. Applications of transliteration editors in the context of Universal Digital Library (UDL) include entry of meta-data and dictionaries for Indian language

The Indian language scripts are syllabic in nature and consist of V, CV, CCV and CCCV type of units, where C is a consonant and V is a vowel. The property of these scripts in that a syllable always ends with a vowel and it makes it easy to identify the syllables using vowels as anchor points.

To render the syllables on the computer screens, Unicode, and the Unicode rendering engine available in Windows XP and Linux operating systems shall be used. To display a CV unit, the UTF-8 sequence of C and V to cause the Unicode rendering engine to render appropriate shape for CV shall be concatenated. To display a CCV unit, we need to render the consonant cluster, so a special character called Halant/Viraam ($) is introduced between every two consonants. So to render CCV we concatenate the UTF-8 sequence of C$CV. To display CCCV type of unit, we concatenate the UTF-8 sequence of C$C$CV. In these syllables, if the vowel is of type schwa (short vowel /a/) then it is nullified as the last consonant in the syllable inherits it by default.

Every consonant in the Indian language scripts inherits schwa and Unicode representation. However, if the vowel is a non-schwa, then the UTF-8 sequence of Maatra of the corresponding vowel is used. A Maatra is a modified shape of a vowel when it is combined with a consonant. Each vowel has only one Maatra.

B. Representation of Indian language scripts

Having known the syllabic nature of Indian language scripts, it is easy to understand the notation followed by the Unicode to represent Indian language Characters.

The following are the principles used by Unicode to represent the Indian language characters:

(1) A Unicode is assigned to each consonant symbol

(2) Each independent vowel is represented by a Unicode.

(3) Each Maatra is also represented by a Unicode.

(4) Viraam is represented by a Unicode number.

C. Unicode rendering recommendations

It should be noted that half-forms of the consonants are not represented by the Unicode. The half-forms are essential to render Aksharas involving consonant clusters. The Unicode recommendation for rendering resolves the issue of half-forms. These rules describe the mapping between Unicode characters and the glyphs in a font. They also describe the combining and ordering of these glyphs. These recommendations are used to build Unicode rendering engines in Windows and Linux operating systems for displaying Unicode characters.

D. Rendering of Aksharas

Akshara is syllabic in nature, and it is rendered by concatenating a sequence of consonant and vowel symbols. In this section, we will examine the rendering of Aksharas of type, CV, CCV and CCCV for the case of English and Hindi.

E. Algorithm of transliteration

1. Take the input as word.
2. Identify each character in string whether it is consonant (c) or vowel (v).
3. If consonant is followed by vowel, then consider it as a single segment.
4. Else if the c is followed by another c (check the next character); go to step 3.
5. Now check the vowel in the substring whether it is "a" or not (e, i, o, u)
6. If it is "a" then go to consonant mapping. If it is not, look up table; then split the word individually
7. Else, go to vowel table and substitute corresponding vowel and append preceding consonant of vowel related maatra

IV. IMPLEMENTATION

A. Modules

In proposed system there are 3 modules

1. Translation of words in dictionary.
2. Transliteration.
3. Translation of sentence.

B. Translation of words in dictionary

In this module translation of the input English word is done based on mapping from the lookup dictionary. Lookup dictionary is the bilingual English-Hindi dictionary, which contains the English words and the corresponding Hindi synonyms. The user enters the source language word and the mapping takes place and the corresponding target words are generated.

C. Transliteration

In this module the source language words are transliterated directly into target language. Transliteration process is performed with the help of Unicode for representing the Devanagari script. The list of Unicode's are provided with the system.

D. Translation of sentence

In this module given input as sentence. Split the sentence into words. If the word is present in dictionary translation is done. If the word is not present in dictionary transliteration is done. Some rules are added into the system, say for example the Subject Verb Object in English is rearranged as Subject Object Verb in Hindi. If a question word appears in the sentence, then it is rearranged to the end of the sentence.

Test case No	Input	Output	Result
1	My name is gayatri	मेरा नाम गयत्रि है	Pass
2	What is your name?	तुम्हारा नाम क्या है?	Pass
3	Who are you?	आप कौन हैं?	Pass
4	swami Vivekananda is great person	स्वामी/गुरूजी विवेकनन्द बडआ व्यक्ति है	Pass
5	This is a pen	यह एक क़लम है	Pass

6	When is your birthday?	तुम्हारा जन्मदिन कब है?	Pass
7	They are good	वे अच्छा हैं	Pass

V. CONCLUSION

The proposed system is a direct translation system along with transliteration. Transliteration is done for named entities through the use of heuristic rules. For translation, bilingual dictionaries are used. Few grammar rules are also used for rearranging the direct translated form of the sentence.

For future work, the translation system can be made to include morphological analysis and morphological generation. It needs to also eliminate the limitations arising due to vowel clusters that occur in the current transliteration system. More grammar rules could be also added to the system to increase the capability of the system. Besides this, new rules could also be included into the system using statistical machine translation techniques

REFERENCES

[1] Cohen, J.M., "Translation", Encyclopedia Americana, vol. 27, 1986.

[2] David Chiang, “Hierarchical Phrase-Based Translation”Computational Linguistics, 33(2):201228, 2007.

[3] Deepa Gupta, Niladri Chatterjee “Study of Divergence for Example Based English-Hindi”.

[4] Ganapathiraju. M., Balakrishnan.M, Balakrishnan. N., Reddy.R, “Om: One Tool for Man Digital Library (ICUDL)” Hangzhou, China. Journal of Zhejiang University SCIENCE, 6A (11):1348-1353.

[5] Jurafsky Daniel and Martin J, “Speech and Language Processing,”

[6] Reinhard Kneser and Hermann Ney, “Improved backing-off for m-gram language modeling” In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Los Alamitos, California, USA. IEEE Computer Society Press, PP-181–184, 1995.

[7] R.M.K. A. Jain,”AnglaHindi: An English to Hindi Machine-Aided Translation System”. In “Proceedings of MT SUMMIT Sinha and IX”, New Orleans, Louisiana, USA.2003.

[8] Simard M and Pierre P,”Bilingual Sentence Alignment: Balancing Robustness and Accuracy “Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-96), pp. 135-144, Montreal, Quebec, Canada,1996.

[9] Shachi Dave, Jignashu Parikh and Pushpak Bhattacharyya, “Interlingua-based English- Hindi Machine Translation and Language Divergence” Journal of Machine Translation, 16 (4): 251-304, 2000.

[10] Sivaji Bandyopadhyay,” State and Role of Machine Translation in India. Machine Translation Review” 2000.