Approaches for multilingual translator for Indian Languages

Harjit Singh
Punjabi University Neighbourhood Campus,
Dehla Seehan (Sangrur), Punjab, India
(E-mail: hjit@live.com)

Abstract**—** India is a multilingual country. Based on languages, the country is divided into states. Even in the same state the language changes over short distances. So Indian literature is available in various languages and even in India the people are not able to understand literature of some other region. IT can be a useful tool to provide NLP to fulfill the gap between languages. NLP is a branch of AI which correlates computer science and linguistics. Basically NLP is a field that provides human computer interaction in a natural language instead of a computer language. The research work in NLP requires deep knowledge of linguistics, statistics and computer science. So it can be categorized as a multidisciplinary research area. NLP can play a very useful role in language conversions such as from Punjabi to Hindi, Gujarati to Punjabi, and Hindi to Gujarati etc. In this way, it can provide access to diverse literature present in regional language to other regions of country. Some Indian languages are easy to convert e.g. from Hindi to Punjabi and vice versa, but some languages are very difficult to convert e.g. from Urdu to Hindi or Punjabi. A multilingual translator can be very useful for Government, businesses and public to access information from different regions of country under a single platform. This paper discusses the approaches that can be used to develop a multilingual translator for Indian Languages using Natural Language Processing.

Keywords**—***Natural Language Processing, NLP, Indian Languages Conversion, Approaches in Language Conversion, NLP Steps.*

I. Introduction

Languages classified as natural languages are the languages spoken by the people. Computer languages are the languages understood by the computers. Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) correlated with linguistics, dedicated to make computers understand natural languages. People use natural languages to communicate among themselves, but to communicate with the computers, human have to learn specific computer language. A language may be English, Hindi, Punjabi, Gujarati etc.; it is a set of symbols and rules. Symbols help people understand the world and are combined together to convey information. Rules are for handling of symbols and they shape the way language is spoken or written.

In India, Hindi is considered as the national language but most of the official and business documents are prepared in English. Hindi is the spoken language and understood by large group of the population. Most of the states use their local language as official language. So in government and legal sector, the translations from one language to another may be required in some cases. In business sector also, the language translations are required according to the targeted audience. Some newspapers are published in multiple languages to target the particular audience. Doing the things manual is very time consuming and cumbersome task, so automation is the best alternative with the help of Natural Language Processing.

Digitizing Indian literature is a huge challenge because of the variety of languages in which the literature is available. To overcome the language barriers, NLP can be very useful tool for language conversion.

II. Historical Review of NLP Systems

In seventeenth century some philosophers (Leibniz, Descartes and others) put some proposals to relate words between languages. But these proposals are theoretical and no actual machine development was did. During mid-1930s, a patent for bilingual dictionary was applied by Georges Artsrouni. A Russian philosopher, Peter Troyanskii also came forward with a bilingual dictionary and a method that deals with grammar between languages. Alan Turing in 1950 published the famous paper “Computing Machinery and Intelligence” in which he proposed a criterion of intelligence that is now called “Turing Test”. An overview of Historical development of Natural Language Processing systems is:

A. Georgetown Experiment

Developed in 1950 by Georgetown and IBM and was able to do automatic translation of more than 60 Russian sentences into English language.

B. STUDENT

Developed by Daniel Bobrow in 1964 and was able to solve high school algebra problems.

C. ELIZA

Developed by Joseph Weizenbaum in 1964 and was a simulation of a Rogerian psychotherapist. It was able to rephrase her response with a few grammar rules.

D. SHRDLU

It was developed in 1970 by Terry Winograd and was able to manipulate blocks of different colors. It was able to receive instructions like “Pick up the green box” or “where is yellow block”. It was able to answer the questions such as “What does the red box contain”. SHRDLU was the system that combined syntax, semantics and reasoning about the real world though natural language understanding. The system was able to handle limited number of sentences and those sentences should be about blocks.

E. LUNAR

The natural language database interface system was LUNAR produced in 1972 with ATNs and Woods’ Procedural Semantics. It was introduced at Second Annual Lunar Science Conference in 1971. The name LUNAR was taken from the database used by the system. Its performance was moderately inspiring.

F. LIFER/LADDER

It was a very impressive NLP system developed in 1978. It was a natural language interface to database about US Navy ships. The semantic grammar was used by the system, so it was very much coupled to its domain. The system used the semantic grammar for various user-friendly features such as the ability to add new dictionary entries, to process incomplete input and to define paraphrases. These features made the system very inspiring.

G. Jabberwacky

It was developed in 1982 by Rollo Carpenter as a chatterbot. Its aim was to simulate human chatting in an entertaining manner.

H. Racter

It was developed in 1983 by William Chamberlain and Thomas Etter as a chatterbot that generated English language prose at random.

I. Watson

It was developed in 2006 by IBM and is a question answering system that defeated the best human players in 2011.

III. Natural Language Processing – Simplified View

Natural Language Processing is performed in four phases. These five phases are interrelated and in reality these rarely occur as sequential and separated phases. These phases are as shown in (Fig. 1):

1. Morphological Processing
2. Syntax Analysis (Parsing)
3. Semantic Analysis
4. Discourse Integration
5. Pragmatic Analysis

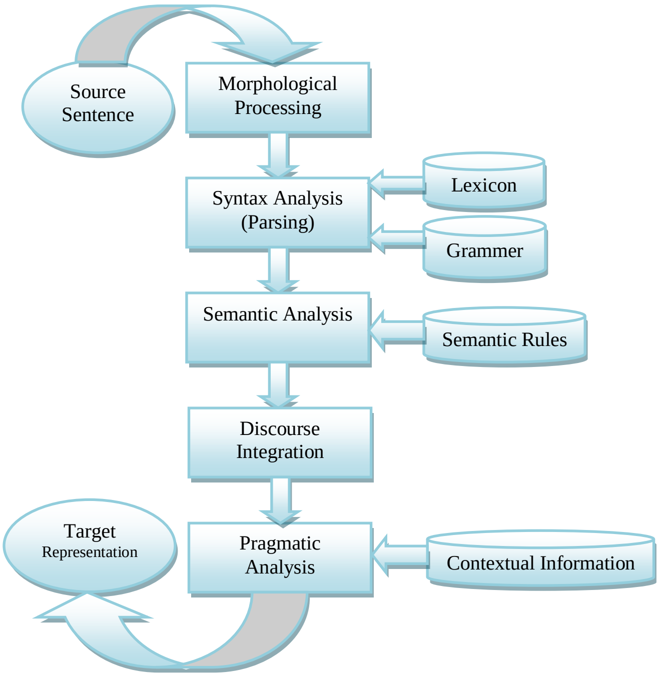

Fig. 1

A. Morphological Processing

The input sentence is composed of tokens and it is decomposed into separate tokens. These tokens can be words, sub-words and punctuation marks. For example, a word such as “decompose” can be broken into sub-words (i.e. tokens) as:

“de" and “compose”

In this phase it is base words are recognized and it is found that how these words are modified to form other words. Words are modified by adding prefixes or postfixes. The phase heavily dependent on the source language being used as input.

B. Syntax Analysis (Parsing)

Syntactic analyzer analyses the format of sentence and checks whether the sentence is well-formed. If so then break it into a specific structure to show the relationship between separate words. The analyzer (called parser) performs its functions by using dictionary (called lexicon) and syntax rules (called grammar).

C. Semantic Analysis

Semantic analyzer needs lexicon and grammar in expanded forms. The lexicon must include semantic definitions of each word and the grammar must specify how semantics sub parts can be used to form semantics of phrases.

D. Discourse Integration

In some sentences, the meaning depends on the preceding sentences. Also it affects the meaning of following sentences. E.g. in the sentence “please have it”, the meaning of “it” depends upon the preceding discourse context.

E. Pragmatic Analysis

Pragmatic analyzer uses the results of semantic analyzer and interprets these results from the viewpoint of a specific context. Sometimes pragmatic analyzer fits actual objects or events that exist in the given context with object references obtained during semantic analysis. The more complicated task of pragmatic analyzer is to disambiguate those sentences which the syntax analyzer and semantic analyzer fail to perform.

IV. INTER LANGUAGE CONVERSION APPROACHES

NLP research for Indian Languages is being done at regional level in the country. But there is a need for multilingual conversion at one place for any Indian language to any other Indian language. There are two approaches that can be adopted to fulfill the need:

A. Conversion using Hindi/English as Intermediate Language

This approach is easy to adopt since the research on regional languages to Hindi/English and vice versa is already being done by NLP researchers in each region or state. All these researches need to be clubbed together to fulfill the need of Government, business and public. Good thing in this approach is that, most of these translators are already available and research is going on to improve their accuracy. So the work done by the researchers can be reused to make if more functional and useful.

As shown in (Fig. 2), It requires only 2xn translators where n is the number of languages we want to integrate. For example, if there are 20 Indian languages we want to integrate in our multilingual translator then 20 translators are required to convert any source language to intermediate language and 20 translators are required to convert intermediate language to any target language. That is, 2x20=40 translators are to be integrated together in multilingual translator.

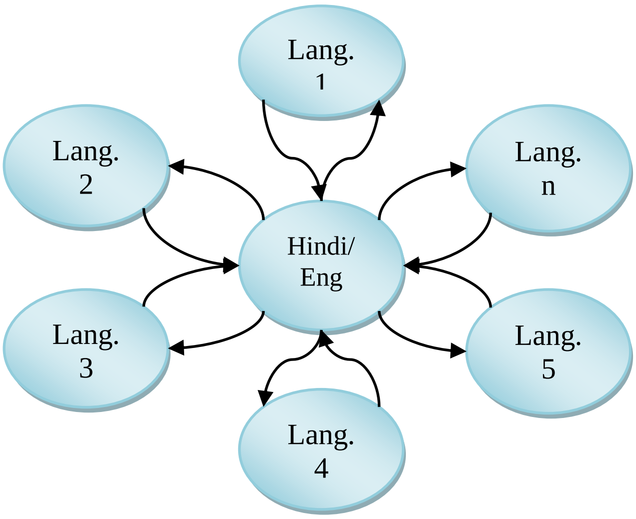

Fig. 2

B. Direct Conversion

The approach discussed above is easy to adopt but it has several disadvantages. First is that, it will consume double time than directly converting source language to target language. Secondly, in such type of conversions, percentage of accuracy may reduce to some or large extent. It is because; the intermediate language used may not be fully compatible with source as well as target language. It means that every language has a specific set of words (vocabulary) and when we convert, we may have to substitute a word that may not be the exact meaning of source word but is near to that one. It is obvious and it may convey the right message in that context but when such word is again converted to target language, it may produce a meaningless result. So the things may become more complicated. (Fig. 3) explains this approach.

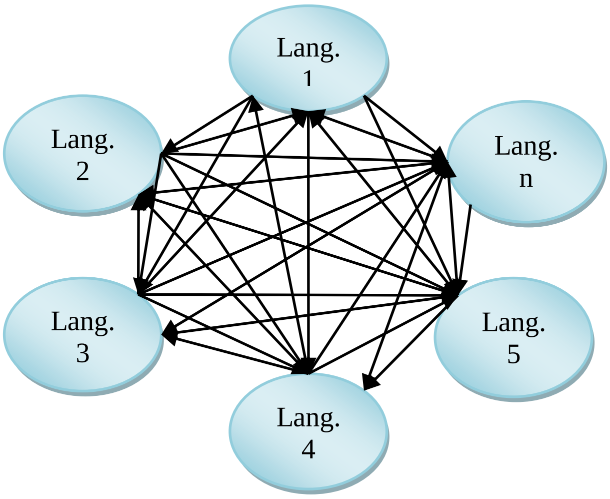

Fig. 3

Alternative but much complicated approach is to convert every source language to every target language directly using separate translator. But it requires large number of translators. It there are n languages then n-1 translators is required for every separate language. That is, nx(n-1) number of translators are needed. For example, if there are 20 languages then 20x19=380 translators are required.

V. CONCLUSION

NLP can play a great role in Indian Language conversions. The research work in language conversion is being done at regional level. Government sector, business sector and even public face difficulties to access information from different regions of country. A multilingual translator can be very useful to fulfill the need.

One easy approach is to use Hindi/English as an intermediate step for conversion since the research on regional languages to Hindi/English and vice versa is already being done by NLP researchers in each region or state. It requires only 2xn translators where n is the number of languages we want to integrate. But this approach will consume double time than directly converting source language to target language. Secondly, in such type of conversions, percentage of accuracy may reduce to some or large extent. Converting every source language to every target language directly using separate translator is another alternative. If there are n languages nx(n-1) number of translators are needed.

REFERENCES

[1] Abhimanyu Chopra, Abhinav Prashar, Chandresh Sain, Natural Language Processing, International Journal Of Technology Enhancements And Emerging Engineering Research, Vol 1, Issue 4

[2] Prof. Langote Manojkumar S, Miss Kulkarni Sweta, Miss Mansuri Shabnam, Miss Pawar Ankita and Miss Bhoknal Kishor, Role of NLP in Indian Regional Languages, IBMRD's Journal of Management and Research Volume-3, Issue-2, September 2014

[3] Gore Lata and Patil Nishigandha, English to Hindi-Translation System,Proceedings of Symposium on translation systems strans (2002).

[4] http://www.slideshare.net/jhonrehmat/natural language processing.

[5] Natural Language Processing,www.myreaders.info /html/artificial_intelligence.html.

[6] Natural Language Processing-Computer science and engineering, www.cse.unt.edu/~rada/CSCE5290/Lectures/Intro.ppt

[7] NLP, https://www.coursera.org/course/nlp

[8] NLP, research.microsoft.com/en-us/groups/nlp/

[9] Dash, N S and B B Chaudhuri. "Why do we need to develop corpora in Indian languages", International Conference on SCALLA, Banglore, 2001

[10] Murthy, B K and W R. Despande. Language technology in India: past, present, and the future. In the Proceedings of the SAARC Conference on extending the use of Multilingual and Multimedia Information Technology (EMMIT'98). Pune, India

[11] Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah and Shata-Anuvadak: Tackling Multiway Translation of Indian Languages, LREC 2014, Rekjyavik, Iceland, 26-31 May, 2014

[12] R M K Sinha. " Machine Translation : An Indian Perspective " , Proceedings of the Language Engineering Conference (LEC'02)

[13] Vishal Goyal and Gurpreet Singh Lehal. "Web Based Hindi to PunjabiMachine Translation System", Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 2, May 2010, pg(s):148-151.

[14] Pushpak Bhattacharyya, Natural Language Processing: A Perspective from Computation in Presence of Ambiguity, Resource Constraint and Multilinguality, CSI Journal of Computing, Vol. 1, No. 2, 2012

[15] https://en.wikipedia.org/wiki/History_of_natural_language_processing

Author received the Bachelor Degree degree in Humanities from the Punjab University, Chandigarh, India, in 2002, the MCA (Master in Computer Applications) degree from IGNOU (Indira Gandhi National Open University), New Delhi, India in 2005 and M.Phil.(CS) degree from Global Open University, Nagaland, India in 2009.

In 2006, he joined the Department of Computer Science at Neighbourhood Campus Dehla Seehan of Punjabi University, Patiala, India as a Lecturer, and later on the post was changed to Assistant Professor in Computer Science. In 2012, he was promoted to Assistant Professor (Senior Scale). He is pursuing Ph.D. degree from RIMT University, Mandi Gobindgarh (Punjab). His current research interests include neural language, image processing and steganography