NLP for Indian Inter language Conversions: Challenges and Opportunities

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

***Abstract*—** India is a country having multiple languages. The states in the country are based on languages; the people speak in those regions. 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. Although research is going on in this field but still the solutions produced do not provide satisfactory results. It is due to the diversity of Indian languages and other challenges like unavailability of Natural Language Processing tools, unavailability of annotated corpora, absence of standards, ambiguity in conversion, unmatched word in target languages etc. 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. This paper discusses the challenges being faced by NLP researchers for Indian Language Conversions.

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. MACHINE TRANSLATION INVENTIONS

An NLP system (Fig. 1) typically uses a computer system which takes input in one language, processes the language to convert it into target language and provides output in target language.

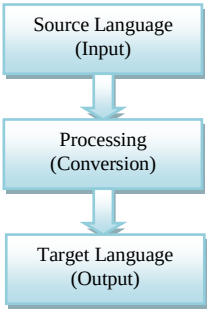

Fig. 1

Various methods are available for machine based translation from one language to another. The output language may not be 100% correct output and so editing need to be done to remove inaccuracies. One such method is Pattern-based Machine Translation, presented by Koichi Takeda from Tokyo Research Laboratory and IBM Research in the proceedings of COLING-96, Copenhagen, Denmark. This method uses a parse tree for conversion process which is structured conversion process. The parse tree of source language sentence is transformed to the corresponding target language tree. Structural conversion can be grammar rule based conversion or template to template conversion.

Another method takes at least one sentence as input and then consults the parsing table for next step. The inventors of this approach include: Duan; Lei (Cupertino, CA), Franz; Alexander M. (Palo Alto, CA) Assignee: Sony Corporation (Tokyo, JP) Sony Electronics, Inc. (Park Ridge, NJ) Appl. No.: 09/240,896 Filed: January 29, 1999. The parser may perform a shift action or a reduce action. The shift action shifts next item from input string into intermediate data structure. Then it generates a new parse node which is associated with a lexical feature. Structure of the shifted input item obtained from a morphological analyzer. This new node is placed in the intermediate data structure. During reduce action; a grammar rule and its associated feature structure are manipulated. If it succeed a new parse node is obtained with the new feature structure. After success, an accept action is performed followed by rebuilding and structural analysis of the input.

In another approach, probabilities or scores are assigned to different target language translations and highest scoring translations are used. The inventors of this approach include: Brown; Peter Fitzhugh (New York, NY), Cocke; John (Bedford, NY), Della Pietra; Stephen Andrew (Pearl River, NY), Della Pietra; Vincent Joseph (Blauvelt, NY), Jelinek; Frederick (Briarcliff Manor, NY), Lai; Jennifer Ceil (Garrison, NY), Mercer; Robert Leroy (Yorktown Heights, NY) Assignee: International Business Machines Corporation (Armonk, NY) Appl. No.: 08/459,454 Filed: June 2, 1995. The source text is converted to intermediate structured representation. These representations are processed to generate intermediate target structure hypotheses. Two different models are used to score these hypotheses. A language model assigns a score to an intermediate target structure. A translation model assigns a score to the source translation event. Both scores are combined to a combined score for every intermediate target structure hypotheses. The highest scoring target structure hypotheses are used to produce target text hypotheses.

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. 2):

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

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.

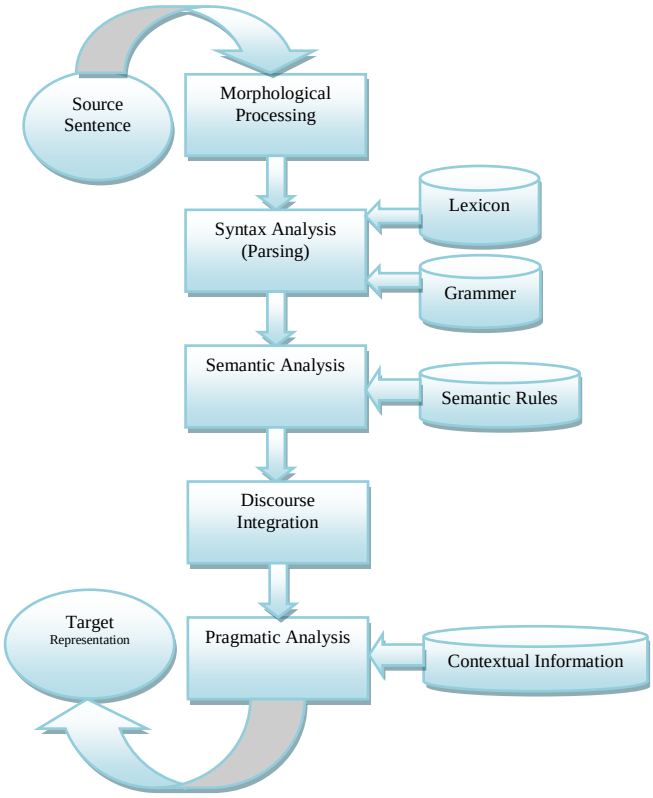

Fig. 2

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. CHALLENGES AND OPPORTUNITIES

NLP research for Indian Languages is being done by researchers at individual level in the country. There are lots of challenges being faced by the researchers in NLP research area:

A. NLP Tools Unavailable

Natural Language Processing tools include dictionaries, lexicons, POS (Part-of-Speech) tagger, morphological generator etc. Unfortunately, these tools are not readily available for Indian Languages. The researchers have to initiate their work from scratch. IIT (Indian Institute of Technology) Bombay has developed Hindi WordNet as well as Marathi WordNet to help researchers. CIIL (Central Institute of Indian Languages) has also initiated efforts in the field.

B. Annotated Corpora Unavailable

Huge collection of machine readable written or spoken structured text is called corpora and the corpora that provide linguistic information is called annotated corpora. Although research is going on but still there is a problem of non-availability of national archive of annotated corpora. It is due to the diversity of Indian languages which required great effort to develop corpora at that level. DOE (Department of Electronics, Govt. of India) in association with CIIL (Central Institute of Indian Languages) has started work in this field and developed corpora for major Indian languages. But still we are far away from the level of corpora of all Indian languages that we need to assist further research in Natural Language Processing.

C. Absence of Standards

Technology requires standards for continuous research and development. In case of Natural Language Processing these standards must be at Font, Script and Input levels. Some of the drafts presented at these levels include:

Font Level: ISFOC (Intelligence based Script Font Code)

Script Level: ISCII (Indian Script Code for Information Interchange) and UNICODE

Input Level: INSCRIPT (Indian Script) phonetic keyboard layout.

But these are not final and fixed standards.

D. Ambiguity in Conversion

Sometimes it becomes difficult to fit a proper word in a sentence since the word may have multiple meanings. During syntactic analysis the ambiguity becomes difficult to overcome. For example, the sentence:

Mother is preparing food and watching TV serial.

In the above sentence the scope of the subject (i.e. Mother) is ambiguous. From machine's perspective it is not clear that if Mother is only preparing food or she is watching TV serial or doing both these activities.

Similarly, the sentence:

I saw a saw which could not saw.

The meaning of the word "saw" is different at different places in the sentence but it becomes ambiguous for the machine to understand the meaning.

In such cases the easiest way is to present a list of alternatives to get user opinion. More research is needed to be done to solve such type of ambiguity during translation.

E. Word Un-matching

Sometimes while translating no proper matching word found in the target language. For example in Punjabi Language the word "Khaadha Peeta" needs much effort to be translate to English because there will be a single word in English and most other languages for these Punjabi word since they have collective meaning "Eat".

Similarly, the Punjabi words "Fer Milaange" can't be directly translated word by word; its meaning is "bye" in English.

Phonetics can be used to convert such words.

F. Testing Difficulty

The researchers made their full efforts to develop better alternative solutions for Indian language conversions using Natural Language Processing. But the absence of tools for Indian Languages makes it very challenging to test these solutions up to the level. Some limited set of sentences are used to test the solutions but the words or sentences that are rarely used in some language remain unchecked that rise to the problem in accuracy of these solutions.

Black box testing of these solutions is an alternative by making the solutions open source. The code can be put on the web so that any number of users familiar with Indian languages can access and use it. Their opinions and suggestions can be accepted for improvement in the developed systems.

V. CONCLUSION

Natural Language Processing 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.

Although research is going on in this field but still the solutions produced do not provide satisfactory results. It is due to the diversity of Indian languages and other challenges like unavailability of Natural Language Processing tools, unavailability of annotated corpora, absence of standards, ambiguity in conversion, unmatched word in target languages etc. So it requires more efforts to make the things better.

The challenges in using Natural Language Processing for Indian languages conversions make the task difficult but not

impossible. The opportunities discussed may provide a gateway to overcome the problems and find better alternatives.

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] Bharati,Akshar,Chaityanya Vineet and Sangal Rajeev,(1995),Natural Language Processing: A Paninian Perspective,Prentice-Hall of India.

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

[5] Cini Kurian, A Review of the Progress of Natural Language Processing in India, International Journal of Advances in Engineering & Technology, Volume 7, Issue 5 (Nov. 2014).

[6] Padariya Nilesh,Chinnakotla Manoj,Nagesh Ajay and Dawant Om P.,(2008),Evaluation of Hindi to English, Marathi to English and English to Hindi.

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

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

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

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

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

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

[13] 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

[14] 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

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

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

[17] 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

[18] 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