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. Each state may have multiple spoken languages. The literature in this country can be found in numerous languages which are so mixed with each other that we may not be even exactly count them. Each region has its own literature and that cannot be understood by some other region. Natural language processing (NLP) is a helping hand in this situation of so many gaps in these languages. NLP makes use of linguistics and the digital field of computer science is related to Artificial Intelligence. NLP is an area that provides a way for interface between a computer and a human through human language. The study in this field requires profound understanding in linguistics with computer science and also statistics. So this research is interdisciplinary in nature. 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. The variation in these languages makes language conversion in some language sets easier. 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 that are spoken by people in their routine communication are called natural languages. The digital machines we call computers do not have the capability to understand these languages. They use computer languages which are not understandable to common people. NLP uses linguistics with its digitization helping hand computer science to make computers capable to get the meaning of natural languages. The computers and humans are separate in the sense that they do not have common set of languages though which they can directly communicate with each other. Each language uses a set of symbolic marks along with a set of rules when used in written.

In India, the country's national language is declared Hindi but commonly almost 100% business and other official documents use English language. On the other hand, the largest population of India is familiar with Hindi and uses it for day to day communication. Since, the states are created based on regional languages, so each state has its own regional language which is also declared as official language of that state. Therefore, on government side as well as in business sector, there is always some need of language to language translation needs. The manual process of these translations is very difficult and in this digital age it is impractical. Automating the translation process is the best practice.

Digitizing the literature also creates a massive challenge due to the diversity of languages used in this literature. This languages related obstacle can be handled with the use of natural language processing.

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.

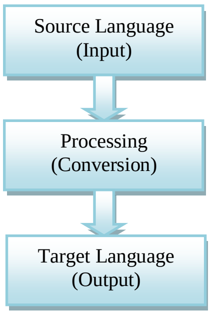

Fig. 1

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. The translation process reads first language text as system-input; apply some set of methods on this input text and produces second language text as output text. A structural conversion method is popular to implement these translation systems in which a structure tree is generated from input words and it is then transformed to another structure tree related to the second language. The methods can adopt a procedure based on grammar rules or from one template to another template conversion.

One such method is based on patterns of text presented by Koichi Takeda. 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 based on grammar rule conversion or based on one template to another 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: Lei Duan and Alexander Franz. 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: Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederick Jelinek, Jennifer Lai, Robert Mercer. 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. HOW NLP WORKS – AN INTERNAL VIEW

NLP has to do a lot of work to get the required output. It works in phases; each phase performs a particular task and may provide its results to another phase. The processing starts when the user inputs textual information for processing. The input text is then cleaned to remove any unwanted symbols from the input text. We can call it noise. Noise is omitted to get the clean and clear text for further processing. The tokenizing is a process by which the text is further divided into individual words commonly called tokens. The text is processed and the words are separated using the existence of white spaces in between the words. The separated words can be stored in a list or array for further processing.

The text is a set of words separated with white spaces. The words may be simple, but at the same time some words may be complex or composed words. These are the words that are made by the combination of two or more words together. The sub-words have their own meaning and when used in composite words, the composite word may have different meaning in different context. The very simple words that cannot be subdivided further are called morphemes. A process commonly named as morphological analysis has the task of finding the morphemes for each word in the text. There are a number of approaches for doing morphological analysis. A common approach is named finite state transducer (FST). The technique is an accepted approach in NLP processing. It uses some data source such as a dictionary with additional information to perform its function successfully. The process takes the individual words as input and presents the stem words and attached modifiers as result. In common cases of natural language processing the stem words are used in further processing ignoring modifiers.

The text is analyzed to get the meaning of each sentence or sub-sentence in the text. The job is performed by a process named syntax analysis also called parsing. This process provides logical meaning of the given text. It requires the sentence should be grammatically correct and the grammar rules are followed for it. The syntax analysis checks for the correctness of the textual sentence. The sentence is put into a structure that conveys the relationships among the component words in the sentence. For this purpose, it needs a dictionary and a set of rules of grammar.

Finding the meaning of a language text is the most important and cumbersome processing task in NLP. This is done using a process named semantic analysis. It is the process which is an important helping hand in NLP to understand the meaning of a natural language sentence. To do the job, all the words in the text are read and each word is assigned with a logical role in the sentence as per the grammar rules. For this purpose, each word's surrounded text is also analyzed. It is basically done to disambiguate the meaning of multiple meaning words. The logical structure of each sentence is analyzed to understand the proper meaning. To do the job perfectly, the process uses expanded forms of related lexicon and also the grammar. The lexicon in expanded form must have each word's semantic definition in it to achieve the goal and grammar must have information about the use of semantic sub-parts.

The sense of the textual context is understood using the process of discourse integration. The meaning of a sentence may be affected by the previous sentences in that context. The coming sentences are also affected. For example, in "it is awesome", the word "it" refers to something that is not clean in this sentence. To find its meaning, the previous sentence or sentences need to be analyzed. This is the task performed by discourse integration process. Many theories are related to discourse processing. On the basis of these theories, some models have been implemented which presents a model of coherence. Some structural theories states that we can use structural units to model a discourse. Some other theories suggest the modeling based on intention or beliefs.

The process of information extraction from given textual data is done by pragmatic analysis. It tries to find the actual meaning of text by using the structure set. It deals with social content and tries to analyze how it affects the interpretation of text. For example, the interpretation of the sentence "pick the bread" must be a request. It should not be interpreted as an order. The outputs of semantic analysis are used by pragmatic analysis and the interpretation is done using a particular context. The semantic analysis provides some object references which are used by pragmatic analysis to make use of actual objects of that given context. This process does the job that the syntax and semantic analysis are not capable to do. The pragmatic analysis is capable to do the job that the other analyzers failed 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, consider the popular 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) 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