(IJRST) 2020, Vol. No. 10, Issue No. I, Jan-Mar **e-ISSN: 2249-0604, p-ISSN: 2454-180X**

EXPLORING THE POSSIBILITY OF APPLYING NATURAL LANGUAGE PROCESSING (NLP) IN INTER MULTI-LINGUAL TRANSLATION OF INDIAN LANGUAGES FOR ENHANCED EASE OF INTEROPERABILITY[1]

Vipul Goyal, Hardik Chaudhary

DOI: 10.37648/ijrst.v10i01.009

Received: 19th January, 2020; Accepted: 29th February, 2020; Published: 26th March, 2020

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. Indian language has multiple kinds of literature that are difficult for another person in different regions to understand their language. This can be a useful tool for filling the gap between two languages with the help of NLP. As we know, NLP is a part of AI, which contains computer science and sentiment or linguistics. So, we can say that NLP is a technique which works as a bridge between humans and computer to fill the gap in computer language. It requires deep knowledge of statistics, computer language, and linguistics. So, it can be placed in the multidisciplinary area. Although research is going on in this field, 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, an 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.

[1] *How to cite the article:* Goyal V., Chaudhary H., Exploring the Possibility of Applying Natural Language Processing (NLP) in Inter Multi-Lingual Translation of Indian Languages for Enhanced Ease of Interoperability, IJRST, Jan-Mar 2020, Vol 10, Issue 1, 42-47, DOI: http://doi.org/10.37648/ijrst.v10i01.009

(IJRST) 2020, Vol. No. 10, Issue No. I, Jan-Mar e-ISSN: 2249-0604, p-ISSN: 2454-180X

I. INTRODUCTION

languages that are spoken by the common people are called natural language. Codes are dialects comprehended by PCs. NLP is a field of Artificial Intelligence (AI) connected with etymology, devoted to causing PCs to comprehend regular dialects. Individuals utilize regular dialects to convey among themselves; however, to talk with PCs, human needs to learn explicit codes. A language might be English, Hindi, Punjabi, Gujarati, and so forth.; it is a lot of images and rules. Images assist individuals with understanding the world and are consolidated together to pass on data. Rules are for dealing with images, and they shape the manner in which language is spoken or composed.

In India, Hindi is considered as the national language, yet the greater part of the authority and business reports are set up in English. Hindi is communicated in language and comprehended by a huge gathering of the populace. The greater part of the states utilizes their neighborhood language as an official language. So in the administration and lawful division, the interpretations starting with one language then onto the next might be required now and again. In the business area, additionally, language interpretations are required by focused on the crowd. A few papers are distributed in different dialects to focus on the specific crowd. Doing the things physically is a very tedious and bulky undertaking, so mechanization is the best option with the assistance of Natural Language Processing.

Digitizing Indian writing is an immense test as a result of the assortment of dialects wherein the writing is accessible. To beat language obstructions, NLP can be a valuable device for language change.

II. MACHINE TRANSLATION INVENTIONS

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

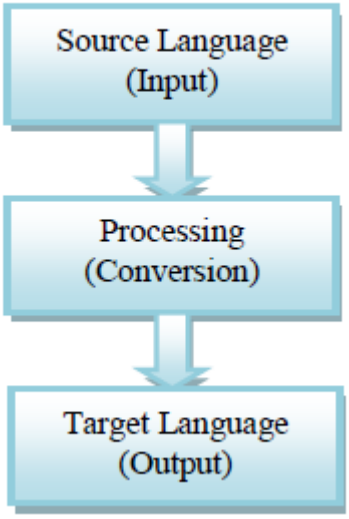

Fig. 1

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 needs 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 the conversion process, which is a structured conversion process. The parse tree of the source language sentence is transformed into 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 the 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 reduce action. The shift action shifts the next item from the input string into an 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 reducing action, a grammar rule and its associated feature structure are manipulated. If it succeeds, a new parse node is obtained with the new feature structure. After success, an accepted 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 the 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 hypothesis. 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 is heavily dependent on the source language being used as input.

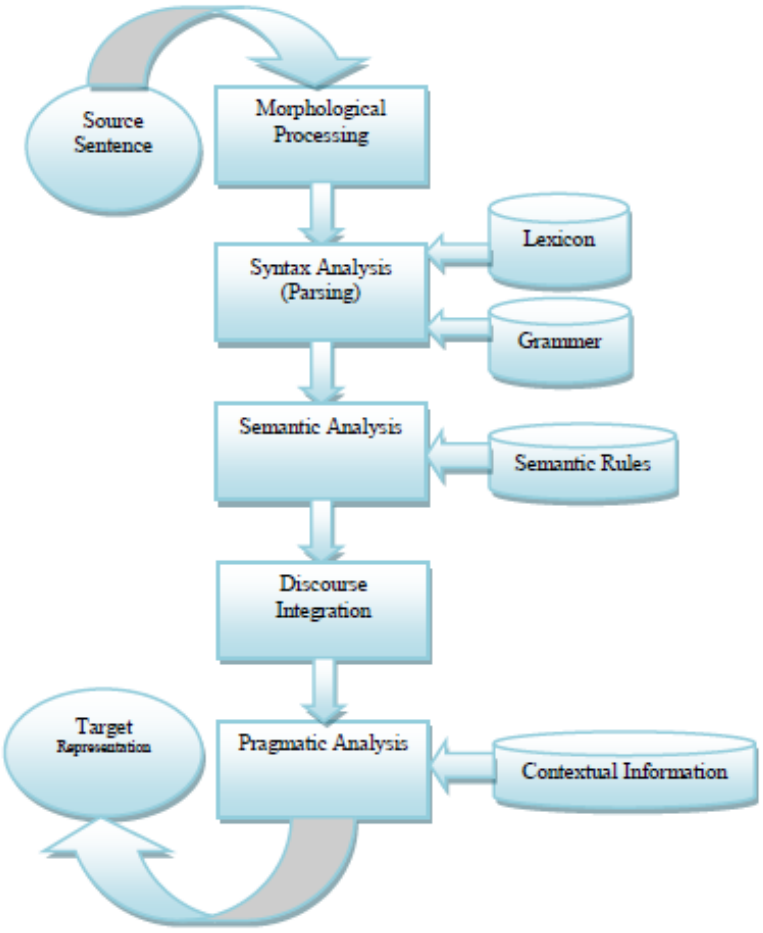

Fig 2

B. Syntax Analysis (Parsing)

Syntactic analyzer analyses the format of the 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 a 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 subparts 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 the following sentences. E.g., in the sentence “please have it,” the meaning of “it” depends upon the preceding discourse context.

E. Pragmatic Analysis

The pragmatic analyzer uses the results of the 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 the 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 an 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

A 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 still, there is a problem of the non-availability of the 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 the 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 serials. In the above sentence, the scope of the subject (i.e., Mother)is ambiguous. From the machine's perspective, it is not clear that if Mother is only preparing food or she is watching TV serials 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 the Punjabi Language, the word "KhaadhaPeeta" needs much effort to be translated into English because there will be a single word in English and most other languages for these Punjabi words since they have collective meaning "Eat." Similarly, the Punjabi words "FerMilaange" 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 inaccuracy 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

NLP can play an extraordinary role in Indian Language transformations. The exploration work in

language transformation is being done at the provincial level. Government division, business area, and even open face challenges are getting to data from various areas of the nation. Although research is going on in this field, 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, the unmatched word in target languages, etc. So it requires more effort to make things better. The challenges in using Natural Language Processing for Indian language conversions make the task difficult but not impossible. The opportunities discussed may provide a gateway to overcome the problems and find better alternatives.