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 affords interaction between human and computer in a language spoken by human. The NLP research and analysis needs good understanding of computer science, statistics and linguistics. So NLP research is a research area having multiple disciplines. NLP will play awfully helpful role in language translations such as from Hindi to Punjabi, Marathi to Punjabi, and Gujarati to Punjabi 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 translation system will be awfully helpful for Business applications, Government agencies and public to approach for information from separate localities of country under one umbrella. This paper discusses the approaches that can be used to develop a multilingual translator for Indian Languages using Natural Language Processing.

Keywords*—*NLP, Translations in Indian Languages, Approach in Language Translation, Steps of NLP.

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.

Hindi is declared as the national language of the country still English is the official and business language in which most of the documents are prepared. Hindi is that the voice communication language and understood by giant cluster of the population. Most provincial offices work in their own local language called language of state. So in legal and government sectors of the country, the conversions from a particular language to another language is needed as per requirements. The language conversions are also needed in business sectors to fulfill the easiness of targeted population. Keeping in view the particular population in the country, some newspaper publishers publish in multiple languages. These bulky tasks are very cumbersome to do manually sentence by sentence, that is why some automation is the need of the time and that automation can be provided by Natural Language Processing.

Now a day, there is a need as well as a trend in digitizing the literature in the country which raises a giant challenge since literature is available in multiple languages. Natural Language Processing can prove itself very helpful in overcoming the language restrictions.

II. HISTORICAL REVIEW OF NLP SYSTEMS

Before the invention of digital computer, there were some proposals about mechanical translators of languages. But the very first NLP system recognized was developed in 1948 at Birkbeck College, London. Warren Weave during Second World War was involved in code-breaking in which a document written in code could be output in another language if code is broken. The concept attracted the research groups. In the beginning, the translation systems developed were for converting German to English. Later on the research expanded to other languages like Russian and French. Earlier systems were not so accurate in producing output. Those systems needed the help of Linguistics to get the required accuracy.

US Funding was stopped for NLP system development after the Automatic Language Processing Advisory Committee (ALPAC) report which concluded the negative view of NLP Research. After that there is much less NLP research work however there were a number of significant developments. Some basic inventions were:

A. Augmented Transition Network

Augmented Transition Network (ATN) is a powerful syntax processing system the make use of grammars. It was not just a syntax processing system instead it is a very powerful

searching software. The system proved itself a very powerful building block in NLP research area where it can produce parses of English sentences.

B. *Case* Grammar

It is related to semantics. English like languages make use of prepositions to express the relationship among nouns and verbs. Charles Fillmore described that many human languages do not use prepositions; still they encode the same type of meaning. Some languages make use of strict word order. Fillmore described that there are very less number of cases where the possible relations among a noun and verb are represented. Individual languages express these relationships using a variety of ways, such as prepositions, word order, word inflection (i.e. the endings of words are changed). It contributed in NLP research in a way that allowed relatively easy theory of implementation such as processing semantic information with little effort.

C. Semantic representations

Conceptual Dependency theory was introduced by Schank and his workers, which is a way of representing language using semantic primitives. Some systems were developed without processing syntax. Quillian's research described the scheme of semantic network, which is awfully helpful for knowledge representation in a number of systems. William Woods described the scheme of procedural semantics for intermediate representation among a database system and a language processing system.

And some popular systems developed were:

D. SHRDLU

SHRDLU system was developed by Terry Winograd which was able to manipulate blocks on a table. It was able to understand commands such as "Pick up the blue pyramid" and was able to answer the queries such as "What does the black box contain?". SHRDLU makes use of combination of semantics, syntax and reasoning to produce a system capable to understand natural language. The systems was very limited to restricted number of sentences and to only the world of blocks.

E. LUNAR

LUNAR was an interface between a database system and a common user and was developed based on ATNs and Woods' Procedural Semantics. The name LUNAR is the name of the database of information about lunar rock samples. The system was introduced in 1971 at the Second Annual Lunar Science Conference. It was able to handle 78% of queries without error.

F. LIFER/LADDER

LIFER/LADDER is a very important development in the area of NLP research. It was developed for the database of information about US Navy Ships to extract information using natural language and proved as a milestone in the NLP research area. It make use of a semantic grammar which made it domain dependent system like SHRDLU. The developers included capability to define new dictionary entries, to process incomplete input and to define paraphrases. These were very impressive features.

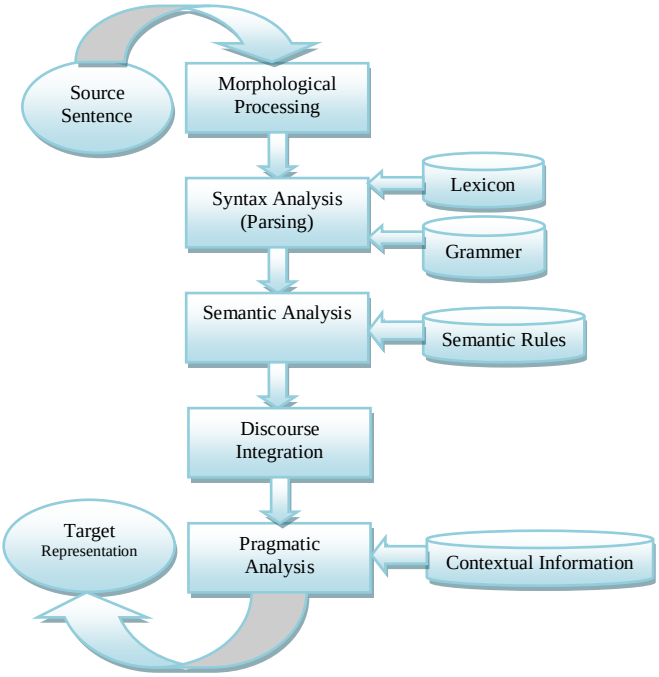

Fig. 1

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

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 APPROACH

The research in Natural Language Processing in India is being performed at regional levels. These efforts are very limited to fulfill individual needs of a particular group of users. Most of the research projects are done at state universities and are not linked or communicated with other universities in the country. It fulfills the requirements of language translations at regional levels but the other part of the country is not taking any advantage of this type of research projects. Most of the research in language translation is done to acquire some educational degree such as Ph.D. and it is mostly for the individual interest of the research scholar or the research guide.

Although a number of language translators are developed by research community related to Natural Language Processing, but they hardly make any concern with other similar research projects being done in other languages. For example, the NLP research in North India is almost unaware of the research in NLP in South India and vice-versa. It may be because of the very difficulty in understanding each others' languages.

The researchers in NLP research community, adopted direct conversion approach in developing language translation systems at regional levels:

A. Direct Conversion

The approach to combine the efforts made by Natural Language Processing researchers and build a multilingual translator from the present translation systems without much additional efforts is just to assemble the developed translation systems into a single multilingual translation system. But it requires some centralized organization to step up so that any difficulties in combining such efforts can be overcome by communicating with each other through that centralized organization. The central government has to establish such organization in a more practical way that it is being done at present.

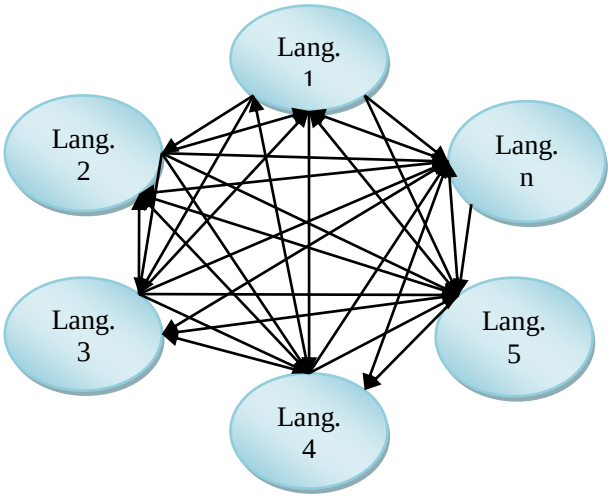

Fig. 2

The approach will make use of separate translator to translate every input language to other output language directly. The translators are already developed at regional levels, the only requirement is to do some modifications in their functioning so that they can be assembled under a single umbrella. The system will be a group of large number of translators. If there are n languages it means that n-1 number of translators are needed for each separate language, most of which may be already available, and others could be developed to fulfill the requirement. That is, nx(n-1) number of translators are needed. For example, if there are 10 languages then 10x19=190 separate translators are needed by the system.

V. CONCLUSION

The research in Natural Language Processing in India is being performed at regional levels. These efforts are very limited to fulfill individual needs of a particular group of users. Other users are unable to get any benefits from this research in NLP. In Legal and Government sectors of the country, the conversions from a particular language to another language are needed. The language conversions are also needed in business sectors to fulfill the easiness of targeted population. A multilingual translator will be awfully helpful to fulfill the need.

The approach is to combine the efforts made by Natural Language Processing researchers and build a multilingual

translator from the present translation systems. The system requires contributions from all the researchers presently developed or developing NLP systems at regional levels plus a centralized organization to assemble these systems together.

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

[16] https://www.cs.bham.ac.uk/~pih/sem1a5/pt1/pt1_history.html