IJCSC

ISSN : 0973-7391

Vol. 3, No. 1, January-June 2012, pp. 193-195

NATURAL LANGUAGE PROCESSING AND ITS MAIN AREAS: INTRODUCTION

Jimmy Singla (AP)[1], Rishideep Singh (AP)[2] and Jagatjit Singh Sekhon (AP)[3]

[1,2,3]North West Institute of Engineering and Technology, Dhudike, Moga, Punjab, India
E-mail: [1]jimmysingla285@gmail.com [2]rishideep80@gmail.com [3]hod_cse@northwest.ac.in

ABSTRACT

This paper makes you familiar with natural language processing and its main areas like speech recognition, speech synthesis, natural language understanding, natural language generation, language translation, need for translation of Indian language, this paper is just an introductory paper from which one can have idea of natural language processing. Natural Language Processing is a technique where machine can become more human and thereby reducing the distance between human being and the machine can be reduced. Therefore in simple sense NLP makes human to communicate with the machine easily.

1. NATURAL LANGUAGE PROCESSING

A natural language is any language used by humans to communicate between them (e.g. English, Hindi, Japanese, and Telugu etc.). Programming languages such as C, C++, and lava etc. are known as artificial languages. Natural language processing refers to descriptions that attempt to make the computers analyze, understand and generate Natural Languages, enabling one to address a computer in a manner as one is addressing a human being. Understanding a human language is not an easy task. The main difficulty lies in knowing the relationship between words, phrases and the concept they represent. A natural language, which is easy for humans to learn and use, is hard for a computer to master. Even long after machines have proven capable of inverting large matrices with speed and grace, they still fail to master the basics of our spoken and written languages. The difficulties in computer processing of a Natural Language from the highly ambiguous nature of natural languages. Very simple sentences for human to speak and understand easily, like "Flying planes can be dangerous", can be very difficult to a computer to understand that lacks knowledge of the world and a native speaker's experience with the linguistic structures of the natural languages. Plausible interpretations of the sentence "Flying planes can be dangerous" could be that "The pilot is at risk", or "There is a danger to people on the ground". Further, should "can" be analyzed as a verb or as noun. Which of the possible interpretations of "plane" is relevant? Depending on the context, "plane" could refer to, among other things, an airplane, a geometric object, or a woodworking tool. How much and what sort of context needs to be brought in to bear on these questions in order to adequately disambiguate the sentence? There are only the few challenges we face while processing a Natural Language.

The term Natural Language Processing represents any processing that is required or need to be done to understand, generate or interpret the utterances in a given language. But in the subsequent paragraphs we list only some main areas or domains of Natural Language Processing:

1.1. Speech Recognition

Speech Recognition is basically the reduction of continuous sound waves to discrete words by computer. Since different people pronounce the same word differently, the mapping of those sounds to the words in the language turns out to be quite difficult.

1.2. Speech Synthesis

Speech Synthesis is the process of generating Natural Language utterances from any type written or handwritten text. As almost every individual's utterance is different, to generate a speech to match the frequency and a style of individuals becomes if not impossible but a very difficult task. The problem of recognizing a written text in itself is quite difficult. For this reason designing and developing systems like Optical Character Recognition (OCR) in itself is a complex task. Thus the synthesis of Natural-Sounding speech is technically complex and almost certainly requires some 'understanding' of what is being spoken to ensure, for example, correct intonation.

1.3. Natural Language Understanding

Natural Language Understanding means moving form words, phrases or sentences (either in written form or

derived by a speech recognition system) to 'meaning'. This involves mapping Natural Language units to their meanings as any Natural Language generates infinite number of valid units; mapping of these dynamically generated units to meaning is quite difficult.

1.4. Natural Language Generation

Natural Language Generation refers to generating appropriate natural language responses to unpredictable inputs.

1.5. Language Translation

Language Translation generally referred as Machine translation (MT) is the application of the computers to the task of translating texts from one natural language to another. One of the very earliest pursuits in computer science, MT has proved to be an elusive goal, but today a number of systems are available that produce output which, if not perfect, is of sufficient quality to be useful in a number of specific domains. Since the present work relates to Machine translation, we will explain in some detail in the following sections.

1.6. Need for Translation of Indian Languages

"India has 18 major regional languages written in 10 different scripts. However, English, though spoken by a minuscule 3 percent of the population, is still the de-facto link language for administration, business and control. All grass root information of land, agriculture, health and education needs to be disseminated in the respective regional languages for effective communication and understanding. Hence, translation is as important as basic and necessary infrastructure like roads, water.

2. TECHNIQUES

There are several main techniques used in analyzing natural language processing. Some of them can be briefly described as follows.

2.1. Pattern Matching

The idea here is an approach to natural language processing is to interpret input utterances as a whole father than builing up their interpretation by combining the structure and meaning of words or other lower level constituents. That means the interpretations are obtained by matching patterns of words against the input utterance. For a deep level of analysis in pattern matching a large number of patterns are required even for a restricted domain. This problem can be ameliorated by hierarchical pattern matching in which the input is gradually canonicalized through pattern matching against subphrases. Another way to reduce the number of patterns is by matching with semantic primitives instead of words.

2.2. Syntactically Driven Parsing

Syntax means ways that words can fit together to form higher level units such as phrases, clauses and sentences. Therefore syntactically driven parsing means interpretation of larger groups of words are built up out of the interpretation of their syntactic constituent words or phrases. In a way this is the opposite of pattern matching as here the interpretation of the input is done as a whole. Syntactic analyses are obtained by application of a grammar that determines what sentences are legal in the language that is being parsed.

2.3. Semantic Grammars

Natural language analysis based on semantic grammar is bit similar to syntactically driven parsing except that in semantic grammar the categories used are defined semantically and syntactically. There here semantic grammar is also involved.

2.4. Case frame Instantiation

Case frame instantiation is one of the major parsing techniques under active research today. The has some very useful computational properties such as its recursive nature and its ability to combine bottom-up recognition of key constituents with top-down instantiation of less structured constituents.

3. APPLICATIONS

Following are the main applications of NLP:

3.1. CoGenTex Inc

CoGen Tex Inc specializes in the development of software systems which represent practical applications of text generation and related areas of natural language processing. These systems developed here help users in various domains to create high quality accurate and maintainable documents either automatically or semi automatically. These systems can be easily be integrated with standard tools such as web browsers and word processors. They also developed software called FoG (Forecast Generator) which generates a textual weather report from a map in both English and French. This has helped many forecasters to produce a weather forecast fast and easily since 1993. Automatic generation frees forecasters from the mechanical aspects of weather reports, allowing them to concentrate on aspects forecasting which most require human knowledge and intuition.

RealPro is CogenTex's cross-platform, high performance syntactic realizer. It is designed to perform syntactic realization, i.e. the transformation of abstract syntactic specifications of natural language sentences (or phrases) to their corresponding surface forms at speed suitable for interactive processing.

Currently CoGenTex is working on several projects focus on diverse topics, from machine translation to computer assisted software engineering.

3.2. NLP at MERL the Enlish Writer's Assistant

Wring text in english presesnts a challenge to non-native speakers because of the difficulties in mastering English vocabulary, grammar and usage. Although most word-processing programs provide some kind of automatic grammar checking, these programmes ae not appropriate for helping non-native speakers to write English text as mistakes they make are different. MERL has developed a system specifically for non-native english speakers and in particular for japanese speakers. The system helps non-native speakers to compose English text while being taught about different aspects of English language has been built. The software developed here although in prototype stage is already very powerful, in addition to correcting the grammar the software demonstrate many useful tool which help the user write Engish text. At MERL they are also working on developing computer supported environment for collaborative learning, with a special focus on constructive and expressive tools for distance learning, work and entertainment.

3.3. Computerise and Online Dictionaries

The idea here is very simple here any word can be looked in the dictionary just by typing and searching for it. This gives fast and very accurate results. Bank of English project by Cobuild is one example.

4. CONCLUSION

Therefore it is clear that Natural Language Processing takes a very important roll in new machine human interfaces. When we look at some of the products that are based on technologies with NLP we can see that they are very advanced but very useful. But there are many limitations, requiring improvements and development of NLP oriented systems. For example language we speak is highly ambiguous. This makes it very difficult understand and analyze. Also with so many languages spoken all over the world it is very difficult to design a system that is 100 % accurate. These problems get more complicated when we think of different people speaking the same language with different styles. Therefore most of research on speech recognition is more concentrated on there areas. Information retrieval can be improved to give very accurate results for various searches. This will involve intelligence to find and sort all the results. So such intelligent systems are being experimented right now are we will be able to see improved applications of NLP in the near future.

REFERENCES

[1] http://en.wikipedia.org/wiki/Natural_language_processing

[2] Katuri Venkateswara Rao, "A Web-Based Simple Sentence Level GB Translator from Hindi to Sanskrit", *School of Compuer and Systems Sciences*, JawaharlaL Nehru.

[3] Joseph Seasly, "Machine Translation: A Survey of Approaches" University of Michigan, Ann Arbor, 2003.