NATURAL LANGUAGE PROCESSING AND RECENT DEVELOPMENTS IN NLP

P SARADA
Department of Mathematics
AMS Arts and Science College for Women,
OU Road
pshardu@gmail.com

P.R. JAYASREE
Department of Mathematics
AMS Arts and Science College for Women,
OU Road
jayasree2395@gmail.com

Abstract:

Perception and communication are essential part of intelligent behaviour. Language is meant for communicating about the world.

A computational model of language, would provide an effective tool for communicating about the world. Language allows for innumerable manners of verbal expression which we can almost always understand. However, that skill is not inherent in machines which require strict parameters in order to function.

The study of Artificial intelligence (AI) allows computer systems to perform tasks that at present require human involvement. Natural language is one of the of the visible forms of AI. Natural Language and machine learning have become the basis for any AI system. In recent times, the field of Natural Language processing has progressed by leaps and bounds. This has allowed computers to understand natural languages.

Natural Language technologies are divided into Natural Language Processing (NLP), which is considered as the umbrella term for the range of natural language technologies, Natural Language Understanding (NLU) regarded as a subset of NLP, Natural Language Interaction (NLI) regarded as converse of NLP and Natural Language Generation (NLG) which is a conflation of above technologies.

In this paper, we explore linguistic facts, especially focusing on the English language. The paper provides detailed information about Natural Language Processing (NLP), steps involved in NLP and its recent developments.

KEYWORDS: *NLP, Language, Syntactic, words, sentences, Knowledge, Program, system, data, AI, words, process, understand, natural, English, tools, Technology*

Introduction:

Humans perceive and communicate through the means of sight and sound to generate meaningful utterances. The two senses involved, are especially complex and require conscious inference. Language is the source of communication.

A computational model of language can be an effective tool for communication. Developing programs that are able to interpret a natural language are difficult to create due to ambiguity in words like “can”, “bear”, ”fly “, “orange” etc which have different meanings in different contexts. Developing Artificial Intelligence (AI) programs with natural forms of communication is essential for user acceptance.

A program understands a Natural Language if it responds to an input with an appropriate action. For example, a student demonstrates understanding if he/she answers a question correctly. A full appreciation of linguistics is not necessarily required in order to understand a natural language, but a familiarity with the basics of grammar is certainly important.

A language understanding program must have significant knowledge about the structure of language including the meaning of different words and how they combine to form phrases and sentences. It must also be able to interpret the context within which a sentence is being used. Here we consider natural language understanding of textual input to information processing. Main three approaches to language understanding are the use of Key word and pattern matching: This is the simplest approach. Matching is used in a variety of programs for different reasons. It may be used to control the sequence and to determine the best no of alternatives and retrieve items from database.

Matching is used in variety of programs like vision, learning, speech recognition automated reasoning programming and planning expert systems and in many ways. Matching is a process which compares two structures and checks equality, for eg.,I prefer to have Pizza, I prefer only Pizza, I want Pizza please. Which is a statement of expression in different ways. In this recognition can be done in by matching key words contained in the statement “Pizza” and negative words are ignored. This analysis contains both structural and semantic analysis.

Parsers are used to analyze individual sentences and to build structures that can be used directly or transformed into required knowledge formats. The advantage of this process is in power and versatility it provides. The disadvantage is that large amount of computation required and need for still further processing to understand the contextual meanings of more than one sentence.

Comparing Matching input to real world situations(Scenario representations) This process is based on structures such as frames or scripts and relies on mapping of input to prescribed primitives which are used to build larger knowledge structures .It depends on the use of constraints imposed by context and world knowledge to develop understanding of knowledge inputs. The advantage is that much of the computation required for computation for syntactical analysis is bypassed. The disadvantage is that a substantial amount of specific as well as general knowledge is prestored. Largest part of Human linguistic communication occurs as speech. Written language place a lesser role than speech in most activities but processing written language is easier than processing speech. Actually, in understanding spoken language we take advantage of clues, such as intonation and the presence of pauses to which we don’t have access when we read. The speech understanding program can be easier enough if know to incorporate into a program knowledge of how to use them. Language provides a facility for the speakers likes to be precise or vague. When speaker knows the hearer’s, language lets the speakers to leave out the things they believe. The Problem that The same expression means different things in different contexts: The source of light is?(From the sun) The source of light is?(From the

Candle) The source of light is?(From the Electricity) and the good side of language is that it allows us to communicate about the world in different ways with limited usage of symbols. Language is to be thought as a program which must have some general world knowledge as well as knowledge of what humans know and how they reason. To get the overall picture we need to think of language as a pair (Source language target representation), together with a mapping between elements of each to the other. In this paper our goal is to be able to reason the knowledge contained in the linguistic expressions and we exploit a frame language as our target representation.

Steps in Process:
1. **Morphological Analysis**-Individual words are analyzed into their components and non-word tokens such as punctuation are separated from the words. Consider the English sentence "I want to pay Susan's. nit file". Morphological analysis is required to do the following things: Pull apart the word "Susan's" into the proper noun "Susan" and the possessive suffix " 's".

Recognize the sequence ".init" as a file extension that functions as an adjective in the sentence also this process will add syntactic category to all the words in the sentence. Interpretations for affixes (prefixes and suffixes).

2. Syntactic Analysis: syntactic analysis shows how words are related to each other by converting the linear sequences of words into structures. Words may be rejected if they violate the language rules. Syntactic analysis must exploit the results of morphological analysis to build a structural description of the sentence. A parsing is a technique which converts flat list of words that forms the sentence into a structure that defines the units that are represented by the flat list. What is important here is a flat sentence has been converted into a Hierarchal structure and that structure has been designed to correspond to sentence units (such as noun phrases) that will correspond to meaning units when semantic analysis is performed. 3.

3. Semantic Analysis: In this analysis meaning are assigned to those structures created by syntactic analyzer. Semantic analysis maps individual words into appropriate objects in the knowledge base or database and how words combine with each other.

Discourse Analysis: The meaning of an individual sentence may depend on the sentences that precede it and may influence the meanings of the sentences that follow it. For eg., the word "It" in the sentence "Gita wanted it" depends on prior discourse context while the word "Gita" may influence the meanings of latter sentence such as "She always had".

5. Pragmatic Analysis: The structure representing what was said is reinterpreted to determine what was actually meant. All the above phases are sometimes performed in sequence and sometimes performed all at once, hence these are often very fuzzy. To make overall understanding tractable we need to decompose a program into the processes and knowledge to perform the task and The global control structure that is imposed on the process.

Natural Language Understanding: In natural language understanding system semantic processing must operate on sentence constituents. If there is no syntactic parsing step then the semantic must decide on its own constituents. If parsing is done on the other hand it constraints number of constituent's semantics can consider. Syntactic parsing is computationally less expensive than semantic processing.

A procedure called parser that compares the grammar against input sentences to produce parsed structures. The most common way to represent grammars is a set of production rules. A sentence is composed of noun phrase followed by verb phrase in this grammar the vertical bar should be read as "OR" the e denotes empty string. Symbols that are further expanded are called non-terminal symbols. Symbols that are found in input sentence are called terminal symbols. Grammar formalisms which underlie many linguistics theories provide a basis for many natural language understanding systems. Regardless of the theory of the basis of grammar the parsing process takes the rules of the grammar and compares them against the input sentence. The simplest structure is to build is a parse tree which simply records the rules and how they are matched. S NP VP PN V NP Bill printed the NPI ADJS N e file Why We Need Rules Of Syntax The Words "boy" "frog" and "eat" occur in a different order in each of the following sentences yet the underlying relation between these words stay the same. The boy ate the frog. The frog was eaten by the boy. The frog which the boy ate died. The boy whom the frog was eaten by died. The second of these examples is referred to in English grammar as a passive sentence. Passive sentences are marked by the object of the sentence appearing in the normal spot for the subject. The main verb is preceded by a form of the verb "to be" (I this case the past tense form "was") and the verb itself "eaten" is in a special form called the past participle.

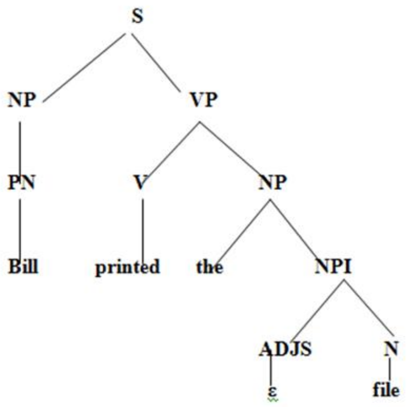

Fig-1

The knowledge used here is knowledge of syntax, a good first thing to do would be to syntactically analyze (or parse) the sentence. Thus, to parse a sentence in, say, English, we need to know the rules of syntax for the language. Then we need some way to apply the rules to actually do the parsing –the parser.

Basic Parsing Techniques

1.**The Lexicon:** A lexicon is a dictionary of words (usually morphemes or root words together with their derivatives), where each word contains some syntactic, semantic, and possibly some pragmatic information.

2.Transition Networks: It is used to represent formal and natural language structures.

They are based on the application of directed graphs (digraphs) and finite state automata. A transition network consists of a number of nodes and labeled arcs. The nodes represent different states in traversing a sentence, and the arcs represent rules or test conditions required to make the transition from one state to the next.

3.Recursive Transition Networks (RTN): Is a network which permits arc labels to refer to other networks and then in turn may refer back to the referring network rather than just permitting word categories used previously.

4. Parsers Deterministic Vs Nondeterministic: A deterministic parser permits only one choice (arc) for each word category. Each arc will have a different test condition. If an incorrect test choice is accepted from some state, the parse will fail since the parser cannot backtrack to an alternative choice. Nondeterministic parsers permit different arcs to be labeled with the same test. Consequently, the next test from any given state may not be uniquely determined by the state and the current input word. The parser must guess at the proper constituent and then backtrack if the guess is later proven to be wrong.

5. Augmented Transition Networks (ATN) the networks considered so far are capable of accepting or rejecting a sentence based on the grammar and syntax of the sentence. To do more useful work an interpreter must be used to build structures and create required knowledge entities for an AI system. The resulting data structures should have more syntactic and semantic information. For additional capabilities RTN (Recursive can be used).

6. An ATN Specification Language: This language is developed by WOODS (1970, 1986) takes the form of extended context free grammar. Using the specification language we can represent this particular network with constituent abbreviation in the form of LISP program.

There are two types of parsing

Top-Down Parsing: It starts with the start symbol and apply the grammar rules forward until the symbols at the terminals of the tree correspond to the components of the sentence being parsed.

Bottom-Up Parsing: Begin with the sentence to be parsed and apply the grammar rules backward until as single tree whose terminals are the words of the sentence and whose top node is the start symbol has been produced.

Few natural language successful understanding languages are LUNAR, LIFER (Language Interface Facility with Ellipses and Recursion) and SHRDLU. LUNAR system is a language interface to give geologists direct access to a database or lunar rock or soil compositions obtained during NASA Apollo-11 moon landing machine.

The main objective was to respond to natural queries such as what is the average concentration of aluminum in high alkali rocks? What is the average of basalt? LIFER (1978) was described by Gary Hendrix. The special features are spelling corrections, processing of elliptical inputs and the ability of the runtime user to extend the language through the use of paraphrase. LIFER has proven to be effective as frontend for a number of systems. The main disadvantage is that due to large number of patterns requires many diverse patterns.

SHRDLU was developed by Terry Win grad (1972, 1986). This system stimulates simple Robot arm that manipulates blocks on a table. The unique aspect of this system is that the meanings of the words and phrases are encoded into procedure that is activated by input sentences. Furthermore, the syntactic and semantic analysis are closely integrated Another way of indicating the same information is in a tree structure as in Figure 2

which expresses the same information, and uses the same terminology, but indicates the internal organization of the sentence by the tree structure, rather than by vertical lines. This notation is widely used in both linguistics and AI. A variant, using slightly different names is shown in Figure-3 Sentence Subject Predicate Verb direct-object Jack ate a frog Tree notation for a diagrammed sentence S NP VP VERB NP JACK ATE AFROG Tree notation using syntactic categories.

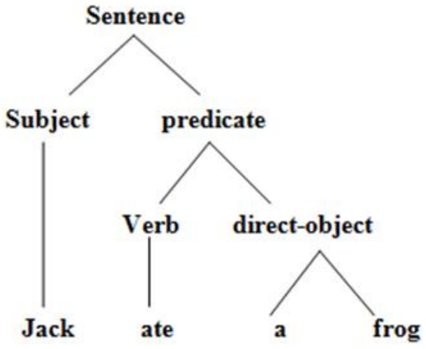

Fig-2 Tree notation for a diagrammed sentence

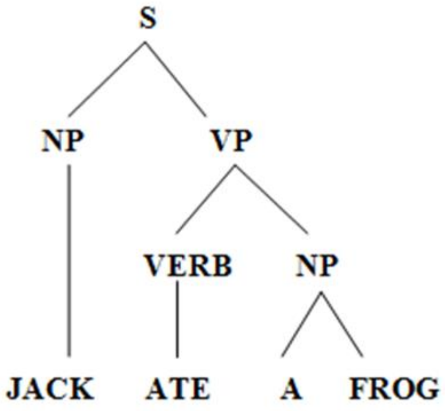

Fig-3 Tree notation using syntactic categories.

Natural Language Generation: This is sometimes claimed as exact inverse of language understanding. The generation of natural language is more difficult than understanding it. Since a system must not only decide what to say but how utterances are to be stated.

A generation must decide which form is better (Active or passive) which words and structures best express the intent and when to say it. To produce expression that are natural and close to human requires more than rules of syntax and discourse that is a coherent plan to be developed to carry out multiple goals and a great sophistication required to convey different shades of meaning and emotions.

A participant in a dialogue must reason out hearers understanding and his/her knowledge and goals. The study of language generation falls into three areas the determination of content: This is concerned with what details include in an explanation, a request, a question or argument in order to convey the meanings set forth by the goals of the speaker.

Formulating and developing a text utterance plan: This is the process of organizing the content to be communicated so as to achieve the goals of the speaker

Achieving a realization of the desired utterances: Realization is a process of mapping the organized content to actual text. This requires that specific words and phrases be chosen and formulated into a syntactic structure. Until about 1980 not much of work was done in NLG.

Recent Developments: The recent developments in NLP employed on machines is user friendly and are growing as per consumer needs due to new inventions of Amazon Alexa, Apple's Siri and Google Home which provide powerful voice recognition and command-based software at very low-price.

However, NLP is by no means perfect because it creates a false positive hit for the words due to change in frequency. To further improve NLP, engineers are constantly developing new programs and implementing revolutionary techniques.

Sentiment analysis is one of the important fields of NLP which detects complex emotions like sarcasm and basic emotions like happiness and sadness. Sentiment analysis is developed by different analytical patterns which provide singular and multiscale patterns. This analysis accurately analyzes text-based inputs.

NLP is taking a lead role whether it is Business Intelligence, Finance, Technology or Health care. NLP techniques are uses in many areas such as sentiment analysis to understand and analyze messages published on social media platforms. NLP with Machine Learning and Big Data is used to resolve queries using Chabot's. Chabot's along with smart assistants like Google Alexa siri and conversational AI allow digital transformation across the world in the areas of sales, marketing, supply chain management, IT help desk and many others. NLP and Big Data are a branch of data science which interpret text similar to human manner and play a key role for analyzing large volumes of text data.

Speech-to-Text: Deep learning is building a path for conversing speech recognition to achieve realistic Human machine interaction. One of the aspects is to build speech to text interfaces which include conversion of audio to text, recognition of audio where the process behind the scene is conversion of Binary (1's and 0's) to text and vice versa and event.

Due to the increasing growth of document flow in organizations more human machine interaction is required in the current business needs and the organizations require to have perspective semantic search for specific terms concepts and business needs which require minute comprehension and Natural language understanding of the core ideas contained in the text achieved with supervised Machine learning techniques.

Cognitive Communication: Cognitive computing is going to be the future of years ahead to tackle various technologies like deep learning, unsupervised and supervised learning though text analytics always plays important role in NLP. In this context NLP falls into distinct categories such as including Natural Language Processing which is focused on linguistics and language's classification and text semantics, Natural Language Understanding, analyzing the actual meaning of words, Natural Language Generation, dealing with the production and generation of text and speech, and Natural Language Interaction a conflation of the above mentioned technologies in which users communicate generate responses from systems through natural language. In this form, NLP is likely to become one of the most visible forms of artificial intelligence in existence

Limitations

Using current knowledge presentation techniques of filtering, re-writing rules, tree structures and matching patterns constraints can be applied on semantic concepts which are enough to narrow the domain.

• The user must adapt to new Machine trends to have effective communicative communication; hand crafting is very much required component and critical facts to be developed for each domain.

• At present systems have limited functionalities of viewing translation, interaction and reading text for sequence processing rather than dealing sentences individually.

References:

[1] Artificial intelligence 2ndedityion Elaine rich Kevin Knight Tata McGraw-Hill publishing co Pvt ltd 1991

[2] Introduction to Artificial Intelligence and Expert systems by Dan w.patterson prentice hall of India Pvt ltd 2003

[3] Introduction to Artificial Intelligence by Eugene Charniak, Drew McDermott published by Pearson Education Pvt ltd 2005

[4] https://aibusiness.com/2019-trends-natural-language-processing/ [5] https://pdfs.semanticscholar.org/b1ab/d494ddae8da5fc388f775f5af748f5cde72f.pdf

[6] https://medium.com/sciforce/natural-language-technologies-in-2019-1bd95a992aa