Natural language processing approaches, application and limitations

Ms. Rijuka pathak
M Tech (CSE) 4 th sem
D.I.M.A.T. Raipur

Mr Biju Thankachan
Associate Profesor C.S.E.
D.I.M.A.T. Raipur

ABSTRACT

Natural language is the language which is used or spoken by the human being these languages are Hindi, English, French, Marathi, Bengali, Gujrati so on. And a natural language processing is the area of artificial intelligence and natural processing is based on creating system through which human can interact with computer in his/her own language without any problem in this paper we focus on natural language processing concept , its approaches , its types and natural language processing application.

INTRODUCTION

Natural language processing is made of three different words 1.natural: natural means not a artificial, which is not created by machine that things are called natural. 2. Language: language is basically a medium of communication through which we make people to understand our thoughts and convey the messages.

As we merge these two word we got natural language as a result so the natural language is the language which is generated automatically not created by machine unlike machine language used in computer ,c , c++ ,java and so on .example of natural languages are Hindi ,English ,Marthi ,Gujrati ,German so on .

And 3rd word is processing the running form of the process which is basically done in computer. So he natural language processing is the processing on natural language through the computer is called as natural language processing (NLP). NLP is basically comes under the area of artificial intelligence basically used for creating language translator, Machine Translation (MT) so on.

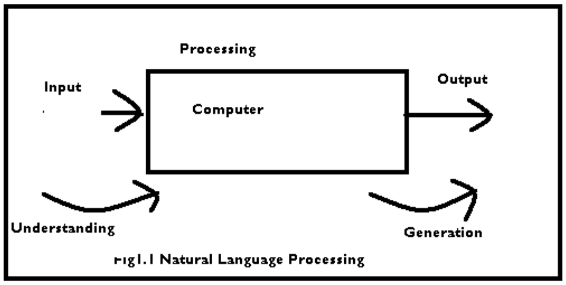


Fig1.1 Natural Language Processing

Fig.1.1 shows the natural language concept NLP process divided into two part understanding and generation .NLP takes

natural language as input understand it process it and generates a desired output

Brief history of NLP

As we know NLP is the subfield of artificial intelligence (AI) and it is new and various developments in NLP are given below:

2.1 First era(1940-1950)

In this period, work mainly confide to two foundational paradigms; namely automation and the use of probabilistic models. the automation started in 1950's with turing's model of algorithmic computations .Turing woke led first to development of MC Culloch pitts neurons, which made a simplified model of neuron as a kind of computing element that could be described in terms of propositional logic ,and then to the work of kleens on finite automation and regular expression. shannon applied probabilistic models of discrete markov processes to automate the language processing. Chomsky, first considered finite state machine a way to characterize grammar and define a finite state language generated by finite state grammar, these model led to field of formal languages theory, which use algebra and set theory to define formal language as sequence symbols. This includes context-free grammar ,first defined by Chomsky.[1]

2.1 Second era (1957-1970)

In second era language processing splits in to two paradigms: symbolic, stochastic. The symbolic work took off from two line of work. the first was the work of Chomsky ,and the other on formal language theory and generative syntax ,and the work of many linguists and computer scientists on parsing algorithms , initially using top-down an bottom-up approaches and then via dynamic programming . The second line of research was the new field of artificial intelligence. These were simple system that worked in single domain mainly by a combination of pattern matching and keyword search with simple heuristics for reasoning and question answering. The decade of 1960 also saw rise of the first testable model of human language processing based on transformation grammar as well as on the online corpora. The brown corpus of American English was a collection of 1 million samples from 500 written texts from different genres which was assembled at Brown University in 1963-64[1].

2.3 Third era (1970-1993)

In this era saw an explosion in research in language processing .the SHRDUL's success showed that parsing was well enough understood to begin to focus on semantics and discourse models. the roger shank and his colleagues built a series of language understanding programs that focused on human conceptual knowledge such as scripts, plans and goals ,and human memory organization

The logic based and natural language understanding paradigms were unified on systems that used predicate logic as a semantic representation. LUNAR is an example of such system.

The discourse-modeling paradigm has been focused on four key discourse analysis.[1]

2.4 Fourth era (1993-till)

In this ea use of probabilistic and data driven models became quite standard throughout natural language processing. Algorithms of parsing, part-of –speech tagging, reference and discourse all began to incorporate probabilities and employ evaluation strategies borrowed from speech recognition and information retrieval.

3. NLP System

As we know NLP has two parts understanding (input side processing) and another part is generation (output side processing). We also know for NLP we need natural language as input and get another natural language as output for this we need general NLP system which follow some steps which is shown in fig no.2 and fig no.3 for input side and another for generation side respectively.

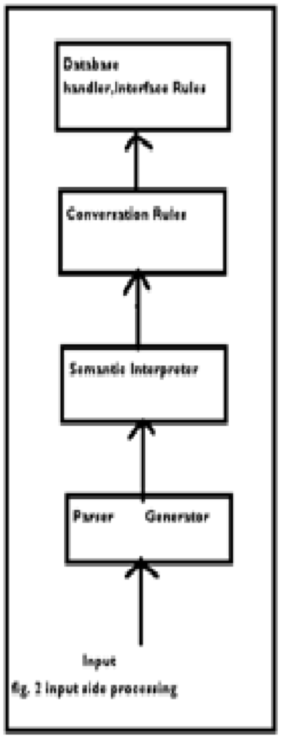


fig. 2 Input side processing

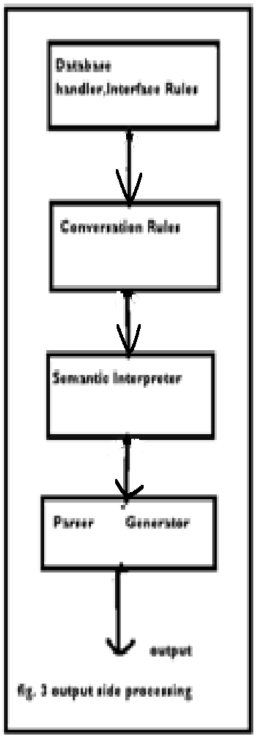


fig. 3 output side processing

Input is given and from here steps are start

1. Parser: Input is given to parser and it generalizes a syntactic structure (in the form of parse tree)

2. Semantic interpreter: then semantic interpreter captures its semantic details generate its deeper structure of it.

3. Conversion Rules: the conversion rules accept this deep structure of sentence and make it compatible for the database storage point of view.

4. Database handler: the database handler works on it generate a processed form from the storage point of view.

And for generation side we need to follow input side processing reverse steps.

4. Linguistic structures

Linguistic is the scientific study of human language. Linguistic can be broadly broke into three categories or subfield of study:

1. Lanugae form

2. Language meaning

3. Langauge in context

Language form: the language form or language structure or grammar .this focus on system rules followed by the speaker of a language. It encompasses- Morphology (the formation and composition of word), syntax (the formation and composition of parse and sentences from these word), and phonology (sound system), and phonetics is a related branch of linguistic concerned with the actual properties of speech sound and non speech sound, and how they are produced and perceived.

Language meaning: the language meaning is concerned with how language employ logical structures and real world reference to convey, process, and assign meaning, as well as resolve ambiguity. This subfield encompasses semantics (how meaning is inferred from words and concepts) and pragmatics (how meaning is inferred from context).

Linguistics in its broader context includes evolutionary linguistics, which considers the origins of language; historical linguistics ,which explores language change, sociolinguistics ,which looks at the relation between linguistic variation and social structure; psycholinguistics, which explores the representation and function of language in the mind ,neurolinguistics ,which looks at language processing in the brain; language acquisition ,how children or adults acquire language ; and discourse analysis; which involves the structure of texts and conversations.

5. Phases of Natural language processing

The natural language processing has six phases- phonology analysis, morphology analysis, lexical analysis, semantic analysis, pragmatic analysis, discourse analysis. All are briefly discussed below-

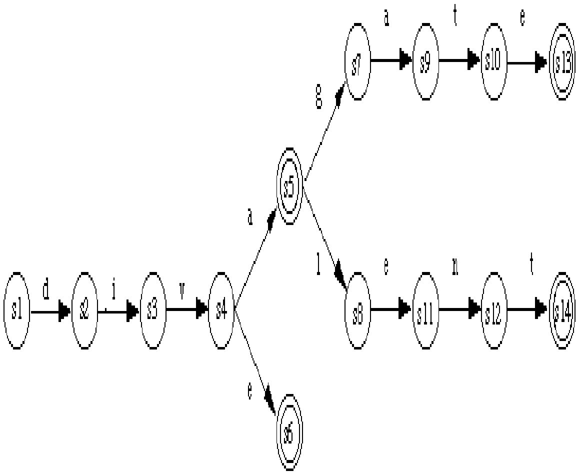
- Phonology analysis: phonology is a branch of linguistics. it is analysis of spoken language .speech recognition and generation
- 1) phonetic rules – for sounds within words;
- 2) phonemic rules – for variations of pronunciation when words ;are spoken together, and;
- 3) Prosodic rules – for fluctuation in stress and intonation across a sentence. In an NLP system that accepts spoken input, the sound waves are analyzed and encoded into a digitized signal for interpretation by various rules.

- Morphology analysis: It is a most elementary phase of NLP. It deals with word formation in this field individual word is analyzed according to there components called “Morphemes”. Morphemes are nothing but basic grammatical building block that makes words. Or in other word structure is referred to as morphology. The computational tools to perform morphological parsing are finite state transducer.

Transducer: A transducer performs it by mapping between the two sets of symbols, and a finite state transducer does it with finite automation.

- A transducer normally consists of four parts:

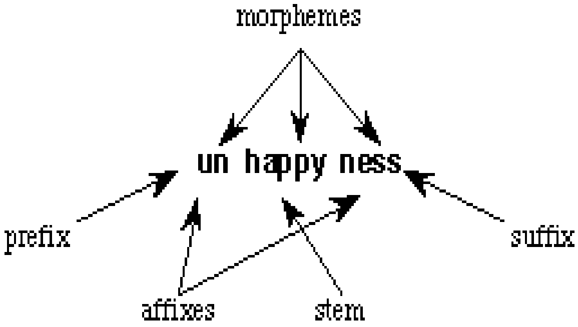
1. recognizer
2. Generator
3. Translator
4. Realtor



Finite state automata

Example of morphology is given below

Example:



- **Lexical analysis: In compute science lexical analysis is the process of converting a sequence of character into sequence of tokens. A program or function which performs lexical analysis is called a lexical analyzer, lexer, or scanner. Lexicon stands for dictionary. It is a collection of all possible valid words of language along with there meaning. Identifying (verb, noun, and pronoun soon).**

Example: consider this expression in the C programming language:

Sum=8+5;

Lexeme	Token Type
sum	Identifier
=	Assignment operator
8	Number
+	Addition Operator
5	Number
;	End of statement

Syntactic analysis

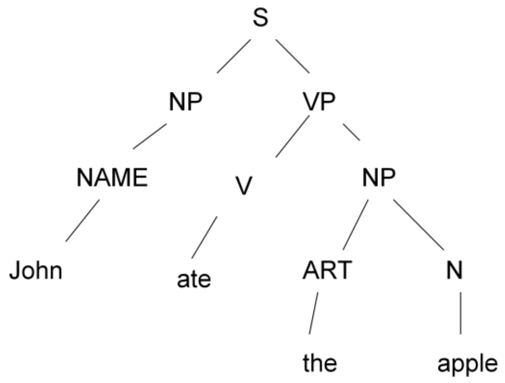
In computer science parsing or more formally syntactic analysis is the process of analyzing a text made of a sequence of token, to determine its grammatical structure with respect to a given formal grammar. The most common grammar used syntactic analysis for natural language are context free grammar also called phrase structure and definite clause grammar.

Types of parser techniques-

1. Top-down parsing

2. bottom-up parsing

e.g.: A parse tree for sentence- john ate apple



s- NP VP

VP- V NP

NP-NAME

NP-ART N

NAME-john

V-ate

ART-the

N-apple

- **Semantic analysis**

Semantic deals with the meaning of natural language sentence. In this phase meaning of sentence is understood. If computer wish to communicate means of

natural language, a computational representation of these sentence is required to capture the meaning.

- Pragmatic analysis: pragmatic is the study of relation between language and context of use .hence pragmatic includes analysis of how language is used to refer to people and things.
- Discourse analysis: the discourse is a collection of sentence, but arbitrary collection of multiple sentences does not make discourse. in discourse the sentence must be coherent

Approaches: Natural language processing approaches fall roughly into four categories: 1.symbolic: symbolic approaches also known as rationalist believe that significant part of the knowledge in human mind is not derived by the senses but is fixed in advance, presumably by genetic inheritance.

2. Statistical

3. Connectionist

4. Hybrid.

Evaluation of NLP:

Intrinsic evaluation: intrinsic evaluation system is based on the concept of measuring the performance of an isolated NLP system and this type and this type of system characterized its performance mainly with respect to a gold standard result, predefined bye the evaluators.

Extrinsic evaluation: It is also known as evaluation in use .it considers the NLP system in a more complex setting either as an embedded or serving is then characterized in terms of its utility with respect to the overall task of the complex system or the human user.

Black box evaluation: black box evaluation requires one to run an NLP system on a given data set and to measure a number of parameter related to quality of the process and most importantly result.

Glass box evaluation: it looks at design of the system, the algorithms that implemented the linguistic resource it uses etc.

Automatic evaluation: automatic procedures can be defined to evaluate an NLP system by comparing its output with the gold standard one.

Manual evaluation: manual evaluation is performed by human judges, which instructed to estimate the quality of a system and most often of a sample of its output, based on a number of criteria.

Application: Natural language processing provides both theory and implementations for a range of

Applications. The most frequent applications utilizing NLP include the following:

- Optical Character Recognition

- Information Retrieval
- Information Extraction (IE)
- Speech recognition

- Text simplification
- Question-Answering
- Summarization
- Machine Translation

• Dialogue Systems

Automatic summarization

• Part-of-speech tagging

• Machine translation

• Named entity recognition

• Natural language generation

• Spoken dialogue system

• Text to speech

.

Limitation: the critical problem with natural language processing is ambiguity refers more then one meaning

Several types of ambiguity are:

Lexical ambiguity: when one word can have several different meanings the resulting ambiguity is called lexical.

Syntactic ambiguity: some times sentences can be parsed in more than one way that is phrases can be put together differently.

Referential ambiguity: the use of pronoun and other anaphora can cause referential ambiguity

Pragmatic ambiguity: this ambiguity underlines meaning of sentence this arises because of different intentions of speaker.

Other are-

Phrase attachment

Conjunction

Noun group structure

Semantic ambiguity

Anaphoric ambiguity

Non-Literal speech

Ellipsis so on.

Conclusion: In throughout the paper we saw the importance of NLP how it help in translation, information retrievals, understanding of different language, machine translation, information extraction soon as NLP has lots of application same time it has some problems also that is ambiguity a lots of research has been done in this field but ambiguity is still a major problem in NLP.

Reference:

[1].Natural language processing ela kumar.

[2]. http://en.wikipedia.org/wiki/Linguistics

[3]. Artificial intelligence by Elaine rich and Kevin knight.

[4].03.nlp.lis.encyclopedia

[5].Natural Language Processing A Paninian Perspective By Akshar Bharti Vineet Chaitanya and Ranjeev Sangal

[6] http://en.wikipedia.org/wiki/Lexical_analysis

[7].http://en.wikipedia.org/wiki/Prosody_(linguistics)

[8]http://en.wikipedia.org/wiki/Semantic_analysis