Contents

Introduction

1

OBJECTIVES

After reading this chapter, the student will be able to Understand:

- Introduction to Natural Language Processing.
- History of NLP .
- Generic NLP Systems .
- Levels of NLP.
- Knowledge in Language processing .
- Ambiguity in NLP .
- Stages in NLP.
- Challenges for NLP.
- Application Areas of NLP:

1. Introduction to Natural Language Processing

Natural Language refers to the language spoken by people, e.g. English, Hindi, Marathi as opposed to artificial/programming languages, like C, C++, Java, etc.

A natural language or ordinary language is any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation. Natural languages can take different forms, such as written text, speech or signing etc. They are distinguished from constructed and formal languages such as those used to program computers or to study logic. Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.

Natural Language Processing (NLP) is a field of research and application that determines the way computers can be used to understand and manage natural language text or speech to do useful things. The term "natural" in the context of the language is used to distinguish human languages (such as Gujarati, English, Spanish and French) from computer languages (such as C, C++, Java and Prolog). The definition of Natural Language Processing clarifies that it is a theoretically induced range of computational techniques (multiple methods or techniques for language analysis) for analyzing and representing naturally occurring text (such as English, Gujarati and Punjabi) at one or more levels of linguistic analysis for the purpose of achieving human like language processing for a range of tasks or applications.

2. Need for Natural Language Processing (NLP)

Significant growth in the volume and variety of data is due to the amount of unstructured text data—in fact, up to 80% of all your data is unstructured text data. Companies collect huge amounts of documents, emails, social media, and other text-based information to get to know their customers better, offer services or market their products .However, most of this data is unused and untouched.

Text analytics, through the use of natural language processing (NLP), holds the key to unlocking the business value within these vast data resources.

In the era of big data, businesses can fully utilize the data potential and take advantage of the latest parallel text analytics and NLP algorithms packaged in a variety of open source software namely R, python etc.

Consider a example given in Figure 1:

kJfmmfj mmmvvv nnnffn333
Uj iheale eleee mnster vensi credur
Baboi oi cestnize
Coovoel2^ ekk; ldsIlk lkdf vnnjfj?
Fgmglmllk mlfin kfre xnnn!

Figure 1: Sample Text in natural form

Computers "see" text in English the same you have seen the figure 1.

Normally, People have no trouble understanding natural language as they have Common sense knowledge, Reasoning capacity, Experience for understanding the context of the text. But this is not the case with Computers. Computers don't have inbuilt Common-sense knowledge, Reasoning capacity, Experience. Unless we teach computers to do so, they will not understand any natural language.

3. Goals of Natural Language Processing:

1. The ultimate goal of natural language processing is for computers to achieve human-like comprehension of texts/languages. When this is achieved, computer systems will be able to understand, draw inferences from, summarize, translate and generate accurate and natural human text and language.
2. The goal of natural language processing is to specify a language comprehension and production theory to such a level of detail that a person is able to write a computer program which can understand and produce natural language.
3. The basic goal of NLP is to accomplish human like language processing. The choice of word "processing" is very deliberate and should not be replaced with "understanding". For although the field of NLP was originally referred to as Natural Language Understanding (NLU), that goal has not yet been accomplished. A full NLU system would be able to:
   - Paraphrase an input text.
   - Translate the text into another language.
   - Answer questions about the contents of the text.
   - Draw inferences from the text.

4. Brief overview of NLP

The field of study that focuses on the interactions between human language and computers is called Natural Language Processing, or NLP for short. It sits at the intersection of computer science, artificial intelligence, and computational linguistics

The essence of Natural Language Processing lies in making computers understand the natural language. That's not an easy task though. Computers can understand the structured form of data like spreadsheets and the tables in the database, but human languages, texts, and voices form an unstructured category of data, and it gets difficult for the computer to understand it, and there arises the need for Natural Language Processing.

There's a lot of natural language data out there in various forms and it would get very easy if computers can understand and process that data. We can train the models in accordance with expected output in different ways. Humans have been writing for thousands of years, there are a lot of literature pieces available, and it would be great if we make computers understand that. But the task is never going to be easy. There are various challenges floating out there like understanding the correct meaning of the sentence, correct Named-Entity Recognition(NER), correct prediction of various parts of speech, coreference resolution(the most challenging thing in my opinion).

Computers can't truly understand the human language. If we feed enough data and train a model properly, it can distinguish and try categorizing various parts of speech(noun, verb, adjective, supporters, etc...) based on previously fed data and experiences. If it encounters a new word it tried making the nearest guess which can be embarrassingly wrong a few times.

It's very difficult for a computer to extract the exact meaning from a sentence. For example – The boy radiated fire like vibes. The boy had a very motivating personality or he actually radiated fire? As you can see over here, parsing English with a computer is going to be complicated.

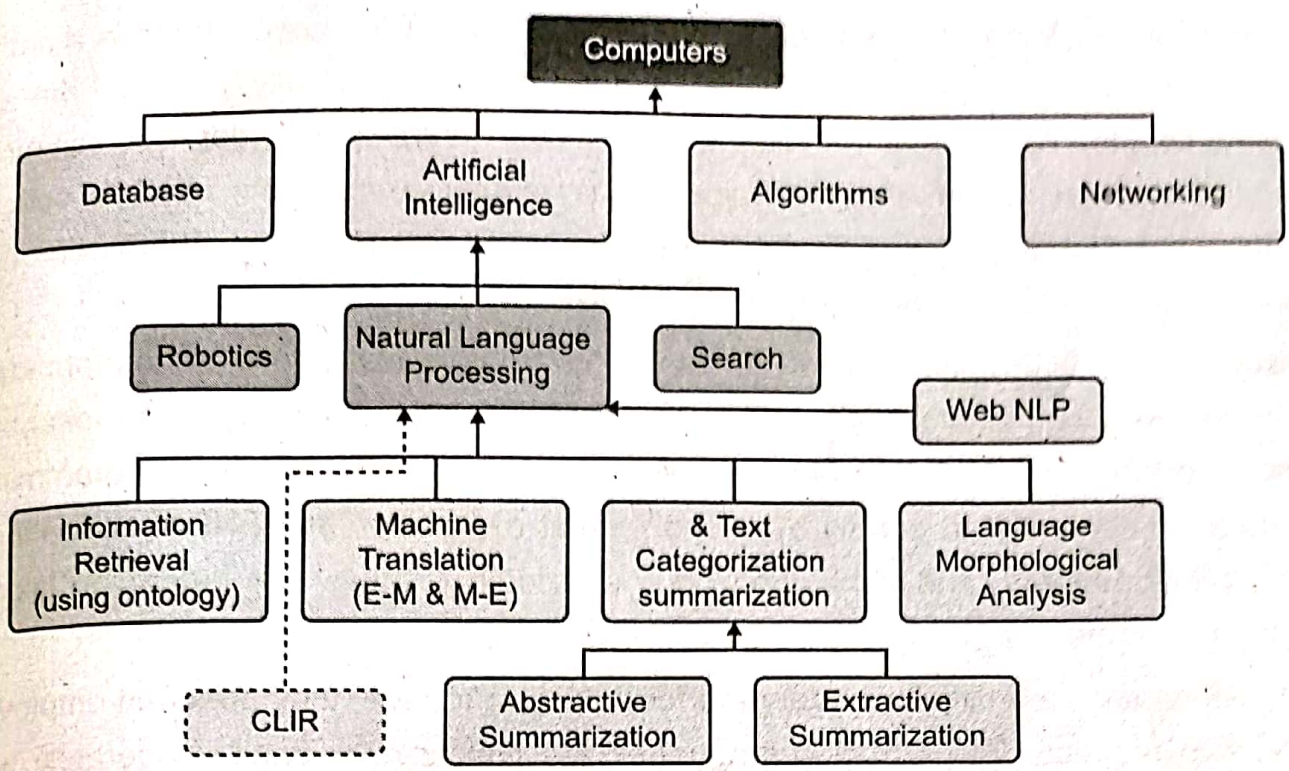


Figure 2: NLP in the Computer science taxonomy

5. History of NLP

NLP began in the 1950s as the intersection of artificial intelligence and linguistics. NLP was originally distinct from text information retrieval (IR), which employs highly scalable statistics-based techniques to index and search large volumes of text efficiently: Manning et al1 provide an excellent introduction to IR. With time, however, NLP and IR have converged somewhat. Currently, NLP borrows from several, very diverse fields, requiring today's NLP researchers and developers to broaden their mental knowledge-base significantly.

Early simplistic approaches, for example, word-for-word Russian-to-English machine translation, were defeated by homographs—identically spelled words with multiple meanings—and metaphor, leading to the apocryphal story of the Biblical, 'the spirit is willing, but the flesh is weak' being translated to 'the vodka is agreeable, but the meat is spoiled.'

Chomsky's 1956 theoretical analysis of language grammars provided an estimate of the problem's difficulty, influencing the creation (1963) of Backus-Naur Form (BNF) notation. BNF is used to specify a 'context-free grammar (CFG), and is commonly used

to represent programming-language syntax. A language's BNF specification is a set of derivation rules that collectively validate program code syntactically. ('Rules' here are absolute constraints, not expert systems' heuristics.) Chomsky also identified still more restrictive 'regular' grammars, the basis of the regular expressions used to specify text-search patterns. Regular expression syntax, defined by Kleene (1956), was first supported by Ken Thompson's grep utility on UNIX.

Subsequently (1970s), lexical-analyzer (lexer) generators and parser generators such as the lex/yacc combination utilized grammars. A lexer transforms text into tokens; a parser validates a token sequence. Lexer/parser generators simplify programming-language implementation greatly by taking regular-expression and BNF specifications, respectively, as input, and generating code and lookup tables that determine lexing/parsing decisions.

While CFGs are theoretically inadequate for natural language, they are often employed for NLP in practice. Programming languages are typically designed deliberately with a restrictive CFG variant, an LALR(1) grammar (LALR, Look-Ahead parser with Left-to-right processing and Rightmost (bottom-up) derivation), to simplify implementation. An LALR(1) parser scans text left-to-right, operates bottom-up (i. e, it builds compound constructs from simpler ones), and uses a look-ahead of a single token to make parsing decisions.

The Prolog language was originally invented (1970) for NLP applications. Its syntax is especially suited for writing grammars, although, in the easiest implementation mode (top-down parsing), rules must be phrased differently (i. e, right-recursively) from those intended for a yacc-style parser. Top-down parsers are easier to implement than bottom-up parsers (they don't need generators), but are much slower.

Recent research has increasingly focused on unsupervised and semi-supervised learning algorithms. Such algorithms are able to learn from data that has not been hand-annotated with the desired answers, or using a combination of annotated and non-annotated data. Generally, this task is much more difficult than supervised learning, and typically produces less accurate results for a given amount of input data. However, there is an enormous amount of non-annotated data available (including, among other things, the entire content of the World Wide Web), which can often make up for the inferior results.

Modern NLP consists of speech recognition, machine learning, machine text reading, and machine translation. These parts when combined would allow for artificial intelligence to gain real knowledge of the world, not just playing chess or moving around an obstacle course. In the near future computers will be able to read all of the information online and learn from it and solve problems and possibly cure diseases. There limit for NLP and AI is humanity, research will not stop until both are at a human level of awareness and understanding.

6. Generic NLP system

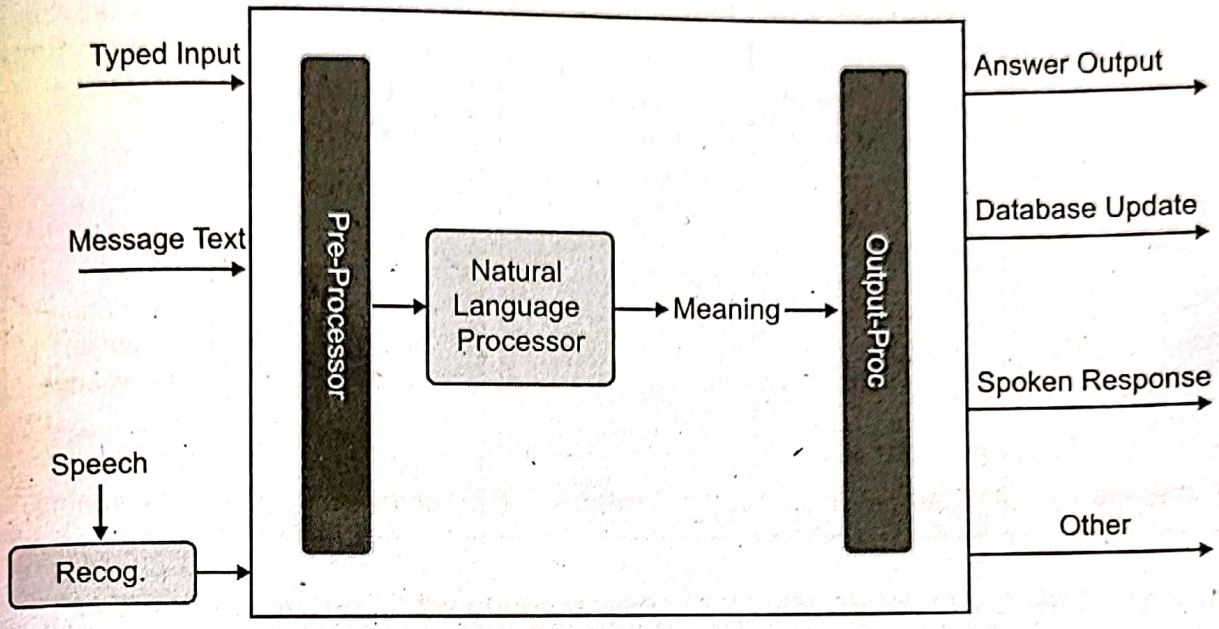
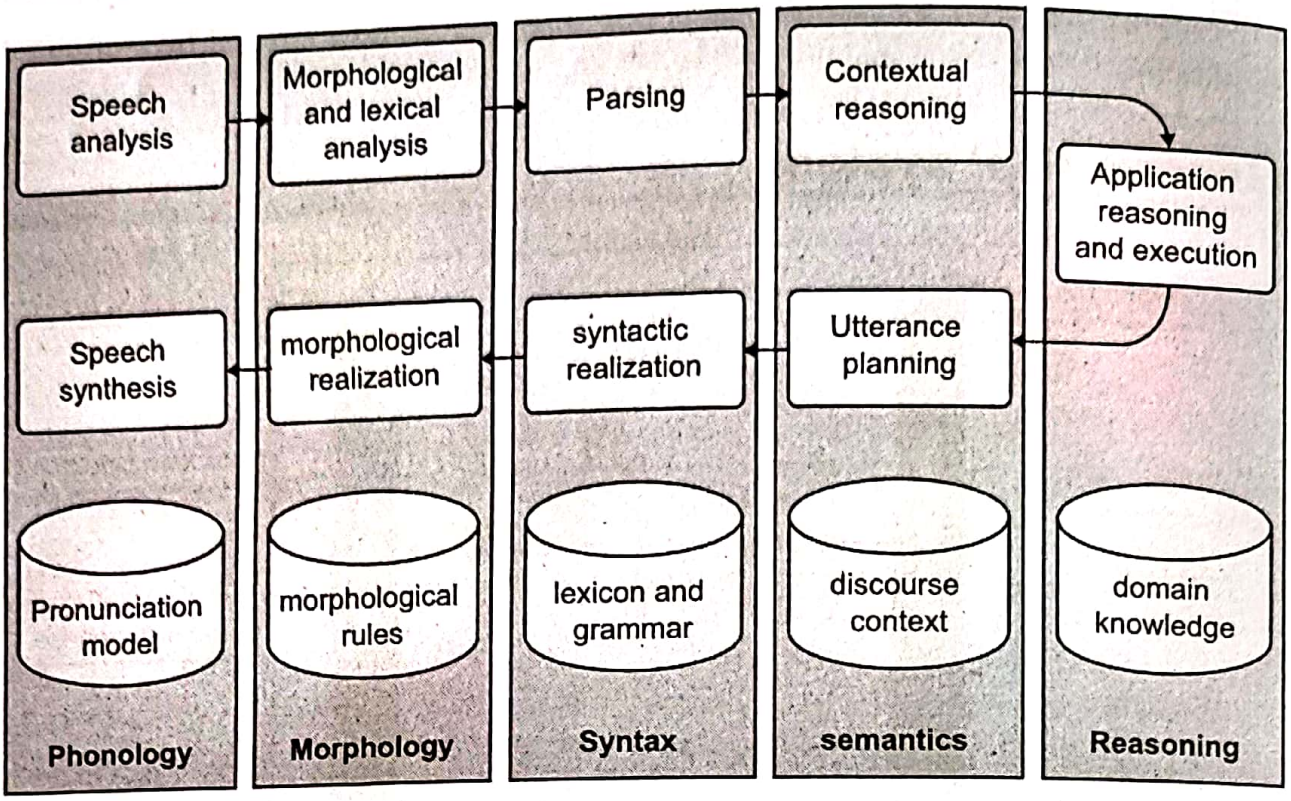


Figure 3: Generic NLP System

Any natural language processing should start with some input and ends with effective and accurate output. The inputs for natural language processor can be text or speech. There are a variety of output that can be generated by the system. Output may be in the form of answer when input is a question. Similarly outputs can be Database update, Spoken response, Semantics, Part of speech, Morphology of word, Semantics of the word/ Sentences etc.

7. Levels of NLP

Natural Language Processing works on multiple levels and most often, these different areas synergize well with each other. The NLP can broadly be divided into various levels as shown in figure.



Phonology: It deals with interpretation of speech sound within and across words.

Morphology: It is a study of the way words are built up from smaller meaning-bearing units called morphemes. For example, the word 'fox' has single morpheme while the word 'cats' have two morphemes 'cat' and morpheme '–s' represents singular and plural concepts.

Morphological lexicon is the list of stem and affixes together with basic information, whether the stem is a Noun stem or a Verb stem [21]. The detailed analysis of this level is discussed in chapter 4. Syntax: It is a study of formal relationships between words. It is a study of: how words are clustered in classes in the form of Part-of-Speech (POS), how they are grouped with their neighbours into phrases, and the way words depend on each other in a sentence.

Semantics: It is a study of the meaning of words that are associated with grammatical structure. It consists of two kinds of approaches: syntax-driven semantic analysis and semantic grammar. The detailed explanation of this level is discussed in chapter 4. In discourse context, the level of NLP works with text longer than a sentence. There are two types of discourse- anaphora resolution and discourse/text structure recognition. Anaphora

resolution is replacing of words such as pronouns. Discourse structure recognition determines the function of sentences in the text which adds meaningful representation of the text.

Reasoning: To produce an answer to a question which is not explicitly stored in a database; Natural Language Interface to Database (NLIDB) carries out reasoning based on data stored in the database. For example, consider a database that holds the academic information about student, and user posed a query such as: 'Which student is likely to fail in Maths subject? To answer the query, NLIDB needs a domain expert to narrow down the reasoning process.

8. Knowledge in Language processing

A natural language understanding system must have knowledge about what the words mean, how words combine to form sentences, how word meanings combine to from sentence meanings and so on. The different forms of knowledge required for natural language understanding are given below.

PHONETIC AND PHONOLOGICAL KNOWLEDGE

Phonetics is the study of language at the level of sounds while phonology is the study of combination of sounds into organized units of speech, the formation of syllables and larger units. Phonetic and phonological knowledge are essential for speech based systems as they deal with how words are related to the sounds that realize them.

MORPHOLOGICAL KNOWLEDGE

Morphology concerns word formation. It is a study of the patterns of formation of words by the combination of sounds into minimal distinctive units of meaning called morphemes. Morphological knowledge concerns how words are constructed from morphemes.

SYNTACTIC KNOWLEDGE

Syntax is the level at which we study how words combine to form phrases, phrases combine to form clauses and clauses join to make sentences. Syntactic analysis concerns sentence formation. It deals with how words can be put together to form correct sentences. It also determines what structural role each word plays in the sentence and what phrases are subparts of what other phrases.

SEMANTIC KNOWLEDGE

It concerns the meanings of the words and sentences. This is the study of context independent meaning that is the meaning a sentence has, no matter in which context it is used. Defining the meaning of a sentence is very difficult due to the ambiguities involved.

PRAGMATIC KNOWLEDGE

Pragmatics is the extension of the meaning or semantics. Pragmatics deals with the contextual aspects of meaning in particular situations. It concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

DISCOURSE KNOWLEDGE

Discourse concerns connected sentences. It is a study of chunks of language which are bigger than a single sentence. Discourse language concerns inter-sentential links that is how the immediately preceding sentences affect the interpretation of the next sentence. Discourse knowledge is important for interpreting pronouns and temporal aspects of the information conveyed.

WORLD KNOWLEDGE

Word knowledge is nothing but everyday knowledge that all speakers share about the world. It includes the general knowledge about the structure of the world and what each language user must know about the other user's beliefs and goals. This essential to make the language understanding much better.

knowledge representation and reasoning systems have incorporated natural language as interfaces to expert systems or knowledge bases that performed tasks separate from natural language processing. As this book shows, however, the computational nature of representation and inference in natural language makes it the ideal model for all tasks in an intelligent computer system. Natural language processing combines the qualitative characteristics of human knowledge processing with a computer's quantitative advantages, allowing for an in-depth, systematic processing of vast amounts of information. The essays in this interdisciplinary book cover a range of implementations and designs, from formal computational models to large-scale natural language processing systems.

9. Ambiguity in NLP

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. The Text based NLP has been regarded as consisting of various levels.

They are:

- Lexical Analysis:- Analysis of word forms
- Syntactic Analysis:-Structure processing
- Semantic Analysis:- Meaning representation
- Discourse Analysis:- Processing of interrelated sentences
- Pragmatic Analysis:-The purposeful use of sentences in situations.

Ambiguity can occur at all these levels. It is a property of linguistic expressions. If an expression (word/phrase/sentence) has more than one interpretation we can refer it as ambiguous.

For e.g: Consider the sentence, **"The chicken is ready to eat".**

The interpretations in the above phrase can be:

- The chicken(bird) is ready to be fed or
- The chicken (food) is ready to be eaten.

Consider another sentence: **"There was not a single man at the party"**

The interpretations in this case can be:

- Lack of bachelors at the party or
- Lack of men altogether

There are different types of ambiguities

1. **Lexical Ambiguity:** is the ambiguity of a single word. A word can be ambiguous with respect to its syntactic class. Eg: book, study.

   For eg: The word "**silver**" can be used as a noun, an adjective, or a verb.

   - She bagged two silver medals.
   - She made a silver speech.
   - His worries had silvered his hair.

   Lexical ambiguity can be resolved by Lexical category disambiguation i.e, parts-of-speech tagging. As many words may belong to more than one lexical category part-of-speech tagging is the process of assigning a part-of-speech or lexical category such as a noun, verb, pronoun, preposition, adverb, adjective etc. to each word in a sentence.

   Lexical Semantic Ambiguity: The type of lexical ambiguity, which occurs when a single word is associated with multiple senses. Eg: bank, pen, fast, bat, cricket etc.

   For eg: 1. The tank was full of water.

   2. I saw a military tank.

Words have multiple meanings for such sentences. Consider the sentence **"I saw a bat."**

Possible meaning of the words which changes the context of the sentence are:

- bat = flying mammal / wooden club?
- saw = past tense of "see" / present tense of "saw" (to cut with a saw.)

The occurrence of tank in both sentences corresponds to the syntactic category noun, but their meanings are different. Lexical Semantic ambiguity resolved using word sense disambiguation (WSD) techniques,where WSD aims at automatically assigning the meaning of the word in the context in a computational manner.

2. **Syntactic Ambiguity:** The structural ambiguities were syntactic ambiguities. Structural ambiguity is of two kinds: Scope Ambiguity and Attachment Ambiguity.

- **Scope Ambiguity:** Scope ambiguity involves operators and quantifiers.

  Consider the example: Old men and women were taken to safe locations.

  The scope of the adjective (i.e., the amount of text it qualifies) is ambiguous. That is,whether the structure (old men and women) or ((old men) and women)?

  The scope of quantifiers is often not clear and creates ambiguity.

  Every man loves a woman

  The interpretations can be, For every man there is a woman and also it can be there is one particular woman who is loved by every man.

- **Attachment Ambiguity**

  A sentence has attachment ambiguity if a constituent fits more than one position in a parse tree. Attachment ambiguity arises from uncertainty of attaching a phrase or clause to a part of a sentence.

  Consider the example:

  The man saw the girl with the telescope.

  It is ambiguous whether the man saw a girl carrying a telescope, or he saw her through his telescope.

  The meaning is dependent on whether the preposition 'with' is attached to the girl or the man.

  Consider the example:

  Buy books for children

  Preposition Phrase 'for children' can be either adverbial and attach to the verb buy or adjectival and attach to the object noun books.

3. **Semantic Ambiguity:** This occurs when the meaning of the words themselves can be misinterpreted. Even after the syntax and the meanings of the individual words have been resolved, there are two ways of reading the sentence.

   Consider the example: **"Seema loves her mother and Sriya does too"**

   The interpretations can be Sriya loves Seema's mother or Sriya likes her own mother.

   Semantic ambiguities born from the fact that generally a computer is not in a position to distinguishing what is logical from what is not.

   Consider the example: **"The car hit the pole while it was moving".**

   The interpretations can be:

   - The car, while moving, hit the pole
   - The car hit the pole while the pole was moving.

   The first interpretation is preferred than the second one because we have a model of the world that helps us to distinguish what is logical (or possible) from what is not. To supply to a computer model of the world is not so easy.

   Consider the example: **"We saw his duck"**

   Duck can refer to the person's bird or to a motion he made.

   Semantic ambiguity happens when a sentence contains an ambiguous word or phrase.

4. **Discourse Ambiguity:** Discourse level processing needs a shared world or shared knowledge and the interpretation is carried out using this context. Anaphoric ambiguity comes under discourse level.

   - **Anaphoric Ambiguity:** Anaphora's are the entities that have been previously introduced into the discourse.

     Consider the example, **The horse ran up the hill. It was very steep. It soon got tired.**

     The anaphoric reference of 'it' in the two situations cause ambiguity. Steep applies to surface hence 'it' can be hill. Tired applies to animate object hence 'it' can be horse.

5. **Pragmatic Ambiguity:** Pragmatic ambiguity refers to a situation where the context of a phrase gives it multiple interpretation. One of the hardest tasks in NLP. The problem involves processing user intention, sentiment, belief world, modals etc. all of which are highly complex tasks.

Consider the example,

"Tourist (checking out of the hotel): Waiter, go upstairs to myroom and see if my sandals are there; do not be late; I have to catch the train in 15 minutes."

Waiter (running upstairs and coming back panting): Yes sir, they are there.

Clearly, the waiter is falling short of the expectation of the tourist, since he does not understand the pragmatics of the situation.

Pragmatic ambiguity arises when the statement is not specific, and the context does not provide the information needed to clarify the statement. Information is missing, and must be inferred. Consider the example: **"I love you too."**

This can be interpreted as:

- I love you (just like you love me)
- I love you (just like someone else does)
- I love you (and I love someone else)
- I love you (as well as liking you)

It is a highly complex task to resolve all these kinds of ambiguities, especially in the upper levels of NLP. The meaning of a word, phrase, or sentence cannot be understood in isolation and contextual knowledge is needed to interpret the meaning, pragmatic and world knowledge is required in higher levels. It is not easy to create a world model for disambiguation tasks. Linguistic tools and lexical resources are needed for the development of disambiguation techniques. Resourceless languages are lagging behind in these fields compared to resourceful languages in implementation of these techniques. Rule based methods are language specific where as stochastic or statistical methods are language independent. Automatic resolution of all these ambiguities contains several long standing problems but again development towards full-fledged disambiguation techniques is required which takes care of all the ambiguities. It is very much necessary for the accurate working of NLP applications such as Machine Translation, Information Retrieval, Question Answering etc.

Statistical Approaches of Ambiguity Resolution in Natural Language Processing are:

1. Probabilistic model
2. Part of Speech Tagging
   - Rule-Based Approaches

- Markov Model Approaches
- Maximum Entropy Approaches
- HMM-Based Taggers

3. Machine Learning Approaches

10. Stages of NLP

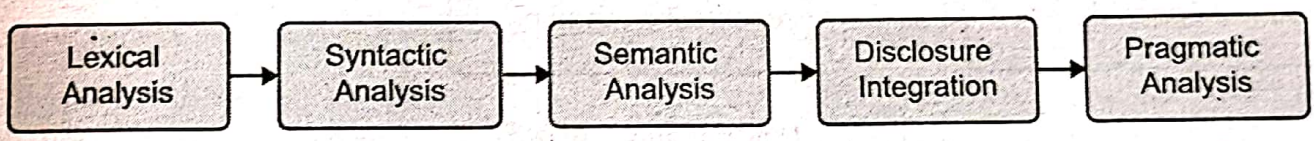
There are general five steps in natural language processing

Lexical Analysis: It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of text into paragraphs, sentences, and words.

The lexical analysis in NLP deals with the study at the level of words with respect to their lexical meaning and part-of-speech. This level of linguistic processing utilizes a language's *lexicon*, which is a collection of individual *lexemes*. A lexeme is a basic unit of lexical meaning; which is an abstract unit of morphological analysis that represents the set of forms or "senses" taken by a single morphemes.

"Duck", for example, can take the form of a noun or a verb but its part-of-speech and lexical meaning can only be derived in context with other words used in the phrase/sentence. This, in fact, is an early step towards a more sophisticated Information Retrieval system where precision is improved through part-of-speech tagging.

Syntactic Analysis (Parsing): It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as "The school goes to boy" is rejected by English syntactic analyzer.



Semantic Analysis: It concerns what words mean and how these meanings combine in sentences to form sentence meanings. It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as "hot ice-cream". Another example can be (plant : industrial plant/ living organism)

Discourse Integration: This concerns how the immediately preceding sentences affect the interpretation of the next sentence. The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of the immediately succeeding sentence.

Pragmatic Analysis: This concerns how sentences are used in different situations and how it affects the interpretation of the sentence. During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

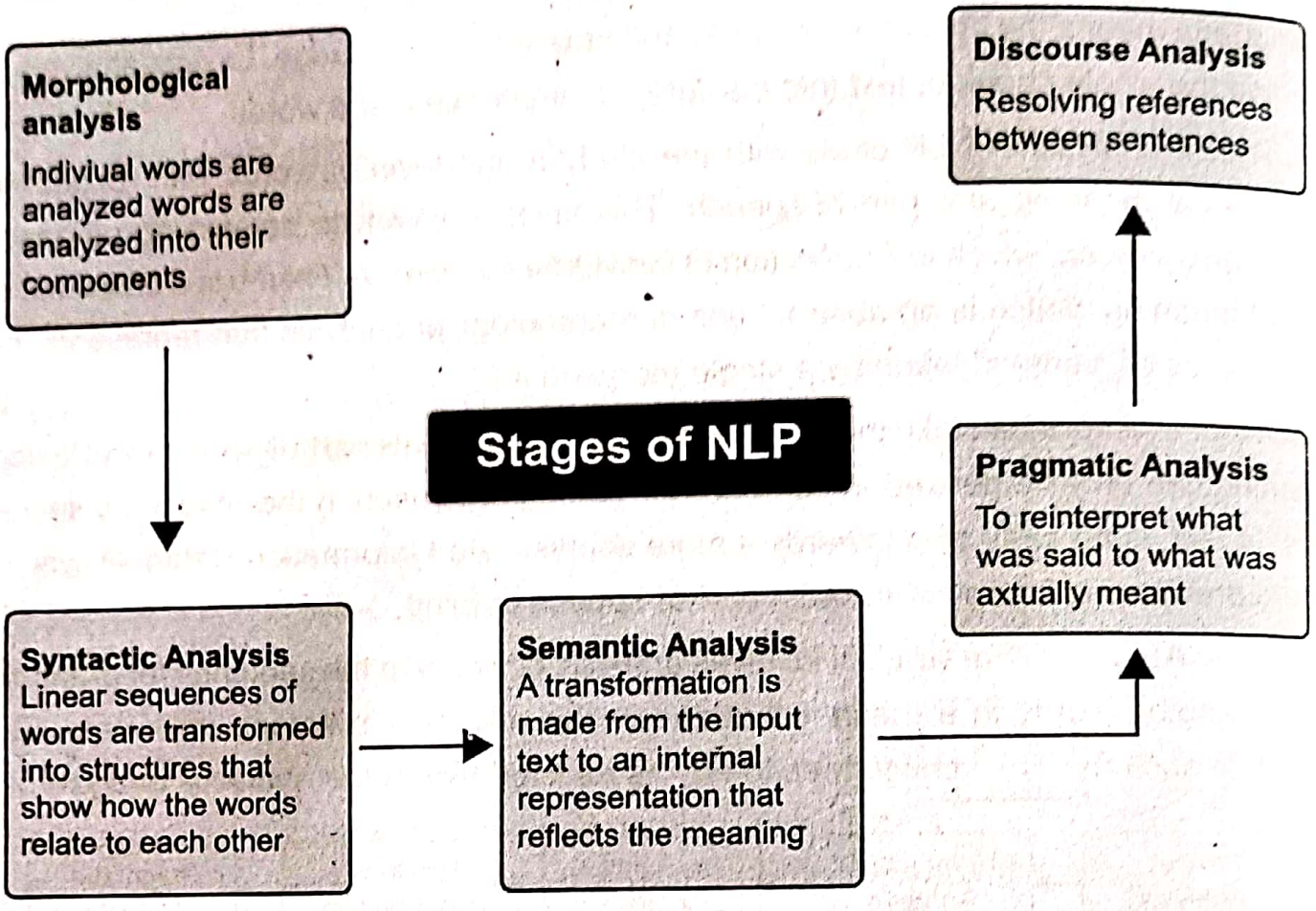


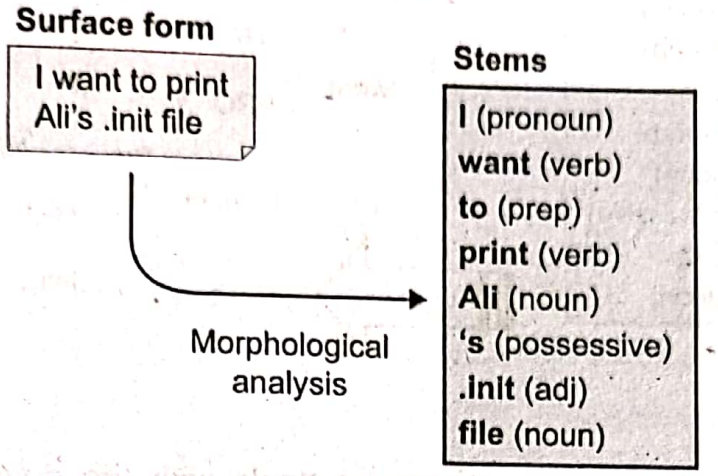
Figure 4: Stages of NLP

Morphological Analysis:

The morphological level of linguistic processing deals with the study of word structures and word formation, focusing on the analysis of the individual components of words. The most important unit of morphology, defined as having the "minimal unit of meaning" is referred to as the *morphemes*. For example, the word: *"unhappiness"*. It can be broken down into three morphemes (prefix, stem, and suffix), with each conveying some form

of meaning: the prefix *un-* refers to "not being", while the suffix *-ness* refers to "a state of being". The stem *happy* is considered as a *free morphemes* since it is a "word" in its own right. *Bound morphemes* (prefixes and suffixes) require a free morphemes to which it can be attached to, and can therefore not appear as a "word" on their own.

In Information Retrieval, document and query terms can be stemmed to match the morphological variants of terms between the documents and query; such that the singular form of a noun in a query will match even with its plural form in the document, and vice versa, thereby increasing recall.



Syntactic Analysis

The part-of-speech tagging output of the lexical analysis can be used at the syntactic level of linguistic processing to group words into phrase and clause brackets. Syntactic Analysis also referred to as *"parsing"*, allows the extraction of phrases which convey more meaning than just the individual words by themselves, such as in a noun phrase.

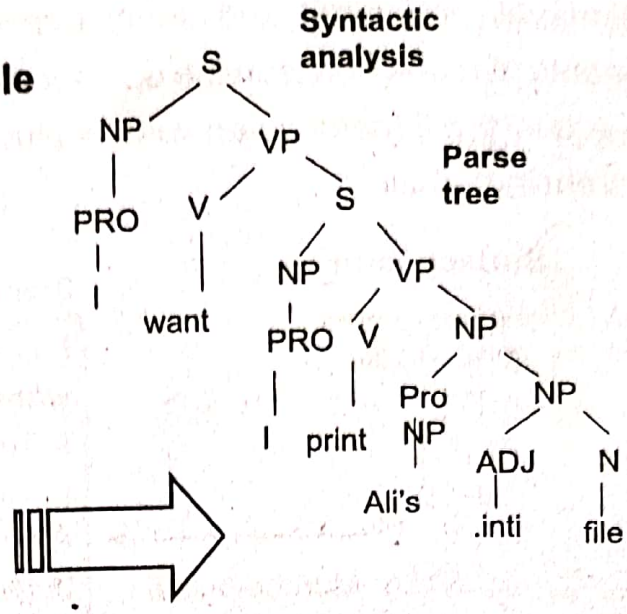
In Information Retrieval, parsing can be leveraged to improve indexing since phrases can be used as representations of documents which provide better information than just single-word indices. In the same way, phrases that are syntactically derived from the query offers better search keys to match with documents that are similarly parsed.

Nevertheless, syntax can still be ambiguous at times as in the case of the news headline: *"Boy paralyzed after tumor fights back to gain black belt"* — which actually refers to how a boy was paralyzed because of a tumor but endured the fight against the disease and ultimately gained a high level of competence in martial arts.

Example

Stems

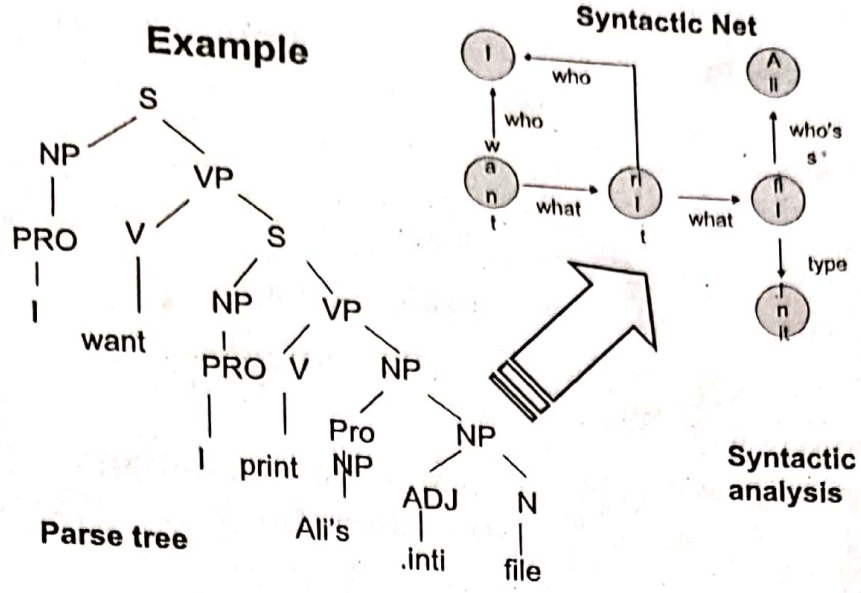
I (pronoun)
want (verb)
to (prep)
print (verb)
All (noun)
's (possessive)
.init (adj)
file (noun)



Semantic Analysis

The semantic level of linguistic processing deals with the determination of what a sentence really means by relating syntactic features and disambiguating words with multiple definitions to the given context. This level entails the appropriate interpretation of the meaning of sentences, rather than the analysis at the level or individual words or phrases.

In Information Retrieval, the query and document matching process can be performed on a conceptual level, as opposed to simple terms, thereby further increasing system precision. Moreover, by applying semantic analysis to the query, term expansion would be possible with the use of lexical sources, offering improved retrieval of the relevant documents even if exact terms are not used in the query. Precision may increase with query expansion, as with recall probably increasing as well.

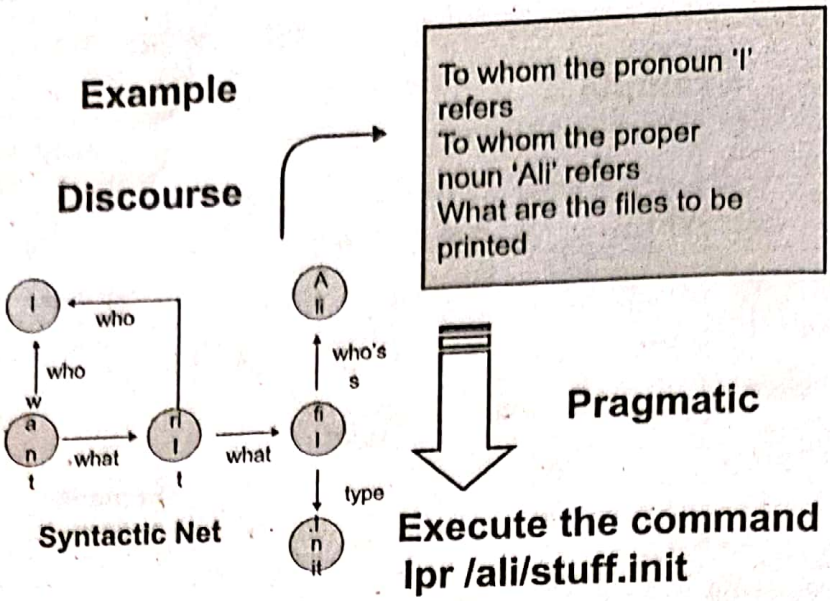


Pragmatic Analysis

The pragmatic level of linguistic processing deals with the use of real-world knowledge and understanding how this impacts the meaning of what is being communicated. By analyzing the contextual dimension of the documents and queries, a more detailed representation is derived.

In Information Retrieval, this level of Natural Language Processing primarily engages query processing and understanding by integrating the user's history and goals as well as the context upon which the query is being made. Contexts may include time and location.

This level of analysis enables major breakthroughs in Information Retrieval as it facilitates the conversation between the IR system and the users, allowing the elicitation of the purpose upon which the information being sought is planned to be used, thereby ensuring that the information retrieval system is fit for purpose.



Discourse Analysis

The discourse level of linguistic processing deals with the analysis of structure and meaning of text beyond a single sentence, making connections between words and sentences. At this level, Anaphora Resolution is also achieved by identifying the entity referenced by an anaphor (most commonly in the form of, but not limited to, a pronoun). An example is shown below.

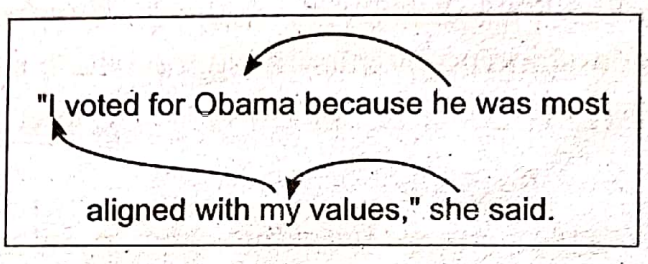


Figure 5: Anaphora Resolution Illustration

With the capability to recognize and resolve anaphora relationships, document and query representations are improved, since, at the lexical level, the implicit presence of concepts is accounted for throughout the document as well as in the query, while at the semantic and discourse levels, an integrated content representation of the documents and queries are generated.

Structured documents also benefit from the analysis at the discourse level since sections can be broken down into (1) title, (2) abstract, (3) introduction, (4) body, (5) results, (6) analysis, (7) conclusion, and (8) references. Information Retrieval systems are significantly improved, as the specific roles of pieces of information are determined as for whether it is a conclusion, an opinion, a prediction, or a fact.

11. Applications of NLP

In the context of Human Computer Interface (HCI), there are many NLP applications such as information retrieval systems, information extraction, machine learning systems, question answering system, dialogue system, email routing, telephone banking, speech recognition system, documentation retrieval system, document summarization, discourse management, multilingual query processing, and natural language interface to database system. Currently interactive applications may be classified into following categories:

Speech Recognition / Speech Understanding and Synthesis / Speech Generation: Speech understanding system attempts to perform a semantic and pragmatic processing of spoken utterance to understand what the user is saying and act on what is being said. The research area in this category includes: linguistic analysis, design & developing efficient and effective algorithms for speech recognition and synthesis.

Language Translator: It is a task of automatically converting one natural language into another preserving the meaning of input text and producing an equivalent text in the output language. The research area in this category includes language modelling.

Information Retrieval (IR): It is a scientific discipline that deals with analysis, design and implementation of a computerized system that addresses representation, organization, and access to large amounts of heterogeneous information encoded in digital format. The search engine is the well known application of IR which accepts query from user and returns the relevant document to user. It returns the document, not the relevant answers; users are left to extract answers from the returned documents. The research area in IR includes: information searching, information extraction, information categorization and information summarization from unstructured information.

Information Extraction: It includes extraction of structured information from unstructured text. It is an activity of filling predefined template from natural language text. The research area in this category includes identifying named entity, resolving anaphora and identifying relationships between entities.

Question Answering (QA): It is passage retrieval in specific domain. It is a process of finding answers for a given question from a large collection of documents.

Natural Language Interface to Database (NLIDB): It is a process of finding answers from database by asking questions in natural language.

Dialog Systems: It is a study of dialog between human and computers. It determines grammar and style of the sentence based on that it gives response to users. The research area in this category includes the design of conventional agent, human-robot dialog and analysis of human-human dialog.

Discourse Management / Story Understanding / Text Generation: The task of identifying the discourse structure is to identify the nature of discourse relationship between sentences such as elaboration, explanation, contrast and also to classify speech acts in a chunk of text (For example, yes-no, statement and assertion).

Expected Questions

1. What is Natural language processing (NLP) ? Discuss various stages involved in NLP process with suitable example.
2. What is Natural Language Understanding? Discuss various levels of analysis under it with example.
3. What do you mean by ambiguity in Natural language? Explain with suitable example. Discuss various ways to resolve ambiguity in NL.
4. What do mean by lexical ambiguity and syntactic ambiguity in Natural language? What are different ways to resolve these ambiguities?
5. List various applications of NLP and discuss any 2 applications in detail.