UNIT- I

Introduction to Natural language

The Study of Language, Applications of NLP, Evaluating Language Understanding Systems, Different Levels of Language Analysis, Representations and Understanding, Organization of Natural language Understanding Systems, Linguistic Background: An outline of English Syntax.

[1]The Study of Language:

- A language is a system of communication which consists of a set of sounds and written symbols which are used by the people of a particular country or region.

Goals

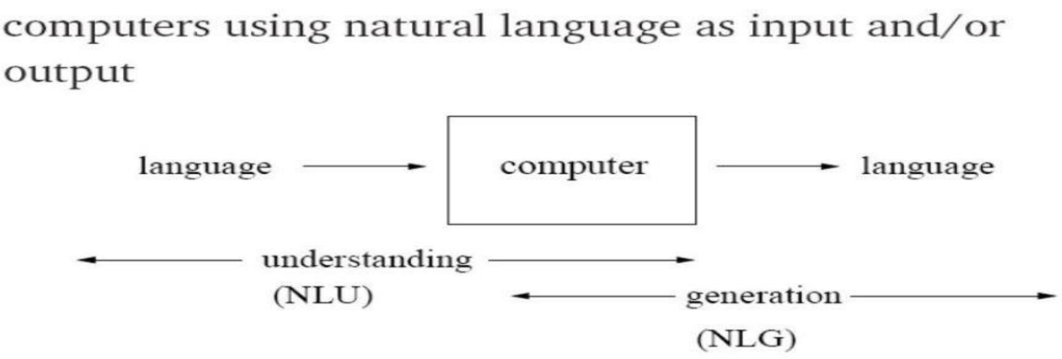
- To create Computational models of language in enough detail that you could write computer programs to perform various tasks involving natural language
- To specify models that approach human performance in the linguistic tasks of reading, writing, hearing, and speaking.

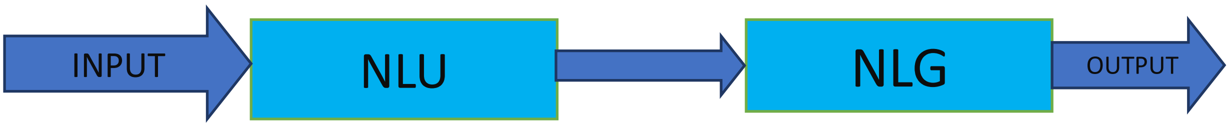
Purpose of computational Models

- Computational models are useful both for
  - Scientific purposes — for exploring the nature of linguistic communication
  - Practical purposes — for enabling effective human-machine communication.

What is Natural Language Processing

- Natural Language process is a branch of computer science and Artificial intelligence used to build the machines that understand and respond to text or voice data and respond with text or speech of their own in the same way humans do.





- ➢ **Natural Language Understanding**
- ➢ **Natural Language Generation**

Natural Language Understanding

What do the users say ? Their intent? Meaning?

Challenges: Lexical Ambiguity
Syntax Ambiguity
Semantic Ambiguity

1. The tank has full of water.
2. Old men and women were taken to safe place.
3. The car hit the pole while it was moving.

Natural Language Understanding

- What should we say to the user
- It should be intelligent and conversational
- It deal with structured data
- Text and sentence planning

Different Academic Disciplines

- **The Linguist**, -Study the structure of language itself.
    - considering questions such as why certain combinations of words form sentences but others do not, and why a sentence can have some meanings but not others.
- **The Psycholinguist-** study the processes of human language production and comprehension.
    - considering questions such as how people identify the appropriate structure of a sentence and when they decide on the appropriate meaning for words
- **The Philosopher** - how words can mean anything at all and how they identify objects in the world.
    - Consider what it means to have **beliefs, goals, and intentions**, and how these cognitive capabilities relate to language
- **The Computational linguist** is to develop a computational theory of language, using the notions of algorithms and data structures from computer science.

Major Disciplines to Study Language

Discipline	Meaning	Typical Problems
Linguists	Structure of language itself	How do words form phrases and sentences? What constrains the possible meanings for a sentence?
Psycholinguists	Processes of human language production and comprehension	How do people identify the structure of sentences? How are word meanings identified? When does understanding take place?
Philosophers	How words can mean anything at all and how they identify objects in the world.	What is meaning, and how do words and sentences acquire it? How do words identify objects in the world?
Computational Linguists	Develop a computational theory of language, using the notions of algorithms and data structures from computer science.	How is the structure of sentences identified? How can knowledge and reasoning be modelled? How can language be used to accomplish specific tasks?

[2]Applications of NLU:

- The applications can be divided into two major classes:
  - **Text-based applications**
  - **Dialogue-based applications.**
- Text-based applications involve the processing of written text, such as books, newspapers, reports, manuals, email messages, and so on. These are all reading-based tasks.

Text-based natural language applications such as

- Finding appropriate documents on certain topics from a data base of texts (for example, finding relevant books in a library)
- Extracting information from messages or articles on certain topics (for example, building a database of all stock transactions described in the news on a given day)
- Translating documents from one language to another (for example, producing automobile repair manuals in many different languages)
- Summarizing texts for certain purposes (for example, producing a 3-page summary of a 1000-page government report)

Dialogue-based applications

- **Dialogue-based applications** involve human-machine communication. Most naturally this involves spoken language, but it also includes interaction using keyboards. Typical potential applications include
- **Question-answering systems**, where natural language is used to query a database (for example, a query system to a personnel database)
- **Automated customer service over the telephone** (for example, to perform banking transactions or order items from a catalogue)
- **Tutoring systems, where the machine interacts with a student** (for example, an automated mathematics tutoring system)
- **Spoken language control of a machine** (for example, voice control of a VCR or computer)
- **General cooperative problem-solving systems** (for example, a system that helps a person plan and schedule freight shipments)

[3]Evaluating Language Understanding Systems:

- To evaluate a system is to run the program and see how well it performs the task it was designed to do.
- If the program is meant to answer questions about a database of facts, you might ask it questions to see how good it is at producing the correct answers.
- **The evaluation has two types 1) Black box evaluation 2) Glass box evaluation**
- If the system is designed to participate in simple conversations on a certain topic, you might try conversing with it.
- This is called **black box evaluation** because it evaluates system performance without looking inside to see how it works.

- Only when the success rates become high, making a practical application feasible, can much significance be given to overall system performance measures.

Glass box evaluation

- This method of evaluation is to identify various subcomponents of a system and then evaluate each one with appropriate tests.
- This is called **glass box evaluation** because you look inside at the structure of the system.
- The problem with glass box evaluation is that it requires some agreement on what the various components of a natural language system should be.
- Achieving such a agreement is an area of considerable activity at the present.

Example: A dialogue with ELIZA

The ELIZA program was developed in the mid-1960s at MIT. This program must be one of the most popular AI

USER	ELIZA
Men are all alike.	In what way?
They're always bugging us about something or other.	Can you think of a specific example?
Well, my boyfriend made me come here.	Your boyfriend made you come here.
He says I'm depressed much of the time.	I am sorry to hear you are depressed.
It's true, I am unhappy.	Do you think coming here will help you not to be unhappy?
I need some help; that much seems certain.	What would it mean to you if you got some help?
Perhaps I could learn to get along with my mother.	Tell me more about your family.
My mother takes care of me.	Who else in your family takes care of you?
My father.	Your father.
You are like my father in some ways.	What resemblance do you see?

- Here is a simple description of how ELIZA works. There is a database of particular words that are called keywords.
- For each keyword, the system stores an integer, a pattern to match against the input, and a specification of the output.
- **The algorithm is as follows:**
  - Given a sentence S.
  - find a keyword in S whose pattern matches S.
  - If there is more than one keyword, pick the one with the highest integer value.

- Use the output specification that is associated with this keyword to generate the next sentence.
- If there are no keywords, generate an innocuous continuation statement, such as "*Tell me more*" or "*Go on*".

Word	Pattern	Outputs
alike	?X	In what way? What resemblance do you see?
are	?X are you ?Y ?X are ?Y	Would you prefer it if I weren't ?Y? What if they were not ?Y?
always	?X	Can you think of a specific example? When? Really, always?
what	?X	Why do you ask? Does that interest you?

Fig: Sample data from ELIZA

- Figure shows a fragment of a database of keywords. In this database a pattern consists of words and variables. **"?X*are you ?Y*"**
- The prefix **?** before a letter indicates a variable, which can match any sequence of words. For example, the pattern would match the sentence ***"Why are you looking at me?",*** where the variable **?X**matches "***Why***" and "**?Y**" matches "***looking at me***".
- The output specification may also use the same variables. In this case, ELIZA inserts the words that match the variables in the input into the output after making some minor changes in the pronouns (for example, replacing "*me*" with "*you*").
- Thus, for the pattern above, if the output specification is "***Would you prefer it if I weren't ?Y?***"
- the rule would generate a response ***"Would you prefer it if I weren't looking at you?"***
- When the database lists multiple output specifications for a given pattern, ELIZA selects a different one each time a keyword rule is used, thereby preventing unnatural repetition in the conversation.
- Using these rules, you can see how ELIZA produced the first two exchanges in the conversation in Figure 1.2.
- ELIZA generated the first response from the first output of the keyword "*alike*" and the second response from the first output of the keyword "*always*".

[4]The Different Levels of Language Analysis:

- A natural language-system must use considerable knowledge about the structure of the language itself, including **what the words are**, **how words combine to form sentences,what the words mean**, **how word meanings contribute to sentence meanings**, and so on.
- However, we cannot completely account for linguistic behavior without also taking into account another aspect of what makes humans intelligent — their general world knowledge and their reasoning abilities.
- For example, to answer questions or to participate in a conversation, a person not only must know a lot about the structure of the language being used, but also must know about the world in general and the conversational setting in particular.

- The following are some of the different forms of knowledge relevant for natural language understanding:
- **Phonetic and phonological knowledge -** concerns **how words are related to the sounds** that realize them. Such knowledge is crucial for speech-based systems.
- **Morphological knowledge** - concerns **how words are constructed from more basic meaning** units called morphemes. A morpheme is the primitive unit of meaning in a language.
- (for example, the meaning of the word "*friendly"* is derivable from the meaning of the noun "*friend"* and the suffix "-*ly"*, which transforms a noun into an adjective).
- **Syntactic knowledge** - concerns **how words can be put together to form correct sentences** and determines what structural role each word plays in the sentence and what phrases are subparts of what other phrases.
- **Semantic knowledge** - concerns what words mean and how these meanings -combine in sentences to form sentence meanings.
- **Pragmatic knowledge** - concerns how sentences are used in different situations and how use affects the interpretation of the sentence.
- **World knowledge** - includes the general knowledge about the structure of the world that language users must have in order to, for example, maintain a conversation. It includes what each language user must know about the other user's beliefs and goals.

Syntax, Semantics, and Pragmatics

- The following examples may help you understand the distinction between syntax, semantics, and pragmatics.

1. Language is one of the fundamental aspects of human behavior and is a crucial component of our lives.
2. Green frogs have large noses.
3. Green ideas have large noses.
4. Large have green ideas nose.

➢ **Sentence 1** appears to be a reasonable start. It agrees with all that is known about syntax, semantics, and pragmatics.

➢ Each of the other sentences violates one or more of these levels.

➢ **Sentence 2** is well-formed syntactically and semantically, but not pragmatically.

➢ **sentence 3** is much worse. Not only is it obviously pragmatically ill-formed, it is also semantically ill-formed.

➢ what is wrong with it: Ideas cannot be green and, even if they could, they certainly cannot have large noses.

➢ **Sentence 4** is even worse. In fact, it is unintelligible, even though it contains the same words as sentence 3. It does not even have enough structure to allow you to say what is wrong with it. Thus it is syntactically ill-formed.

For example, if I ask you **where you are going** and you reply "**I go store**", the response would be understandable even though it is syntactically ill-formed. Thus it is at least pragmatically well-formed.

[5]Representations and Understanding:

- Understanding involves computing a representation of the meaning of sentences and texts.
- For instance, why not simply use the sentence itself as a representation of its meaning?
- One reason is that most words have multiple meanings, which we will call senses.
- The word "**Cook**", has a sense as a verb and a sense as a noun.
- "**Dish**" has multiple senses as a noun as well as a sense as a verb.
- "**Still**" has senses as a noun, verb, adjective, and adverb.

To represent meaning, we must have a more precise language.

- The tools to do this come from mathematics and logic and involve the use of formally specified representation languages. Formal languages are specified from very simple building blocks.
- The most fundamental is the notion of an atomic symbol which is distinguishable from any other atomic symbol simply based on how it is written.

The representation languages have the **following two properties**:

- **The representation must be precise and unambiguous**. You should be able to express every distinct reading of a sentence as a distinct formula in the representation.
- **Therepresentation should capture the intuitive structure of the natural language** sentences that it represents.

Syntax: Representing Sentence Structure

- The syntactic structure of a sentence indicates the way that words in the sentence are related to each other.
- This **structure indicates** how the words are grouped together into phrases, what words modify what other words, and what words are of central importance in the sentence.
- This structure may identify the types of relationships that exist between phrases and can store other information about the particular sentence structure that may be needed for later processing.
- For example, consider the following sentences:

1. John sold the book to Mary.
2. The book was sold to Mary by John.
3. *After it fell in the river, John sold Mary the book.
4. After it fell in the river, the book was sold to Mary by John.

- These sentences share certain structural properties.

• In each, the noun phrases are **"John", "Mary**", and **"the book",** and the act described is some selling action.
• In other respects, these sentences are significantly different.
• For instance, even though both sentences are always either true or false in the exact same situations.
• you could only give **sentence 1** as an answer to the question **"What did John do for Mary?"**
• **Sentence 2** is a much better continuation of a sentence beginning with the phrase
• "After it fell in the river", as sentences 3 and 4 show.

5. *John are in the corner.
6. *John put the book.

• **Sentence 5** is ill-formed because the subject and the verb do not agree in number (the subject is singular and the verb is plural).
• **Sentence 6** is ill-formed because the verb *put* requires some modifier that describes where John put the object.

• **"Making judgments on grammaticality is not a goal in natural language understanding. In fact, a robust system should be able to understand ill-formed sentences whenever possible".**

7. flying planes are dangerous.
8. flying planes is dangerous.
• Consider sentences 7 and 8, which are identical except for the number feature of the main verb, yet represent two quite distinct interpretations.
• If you did not check subject-verb agreement, these two sentences would be indistinguishable and ambiguous.

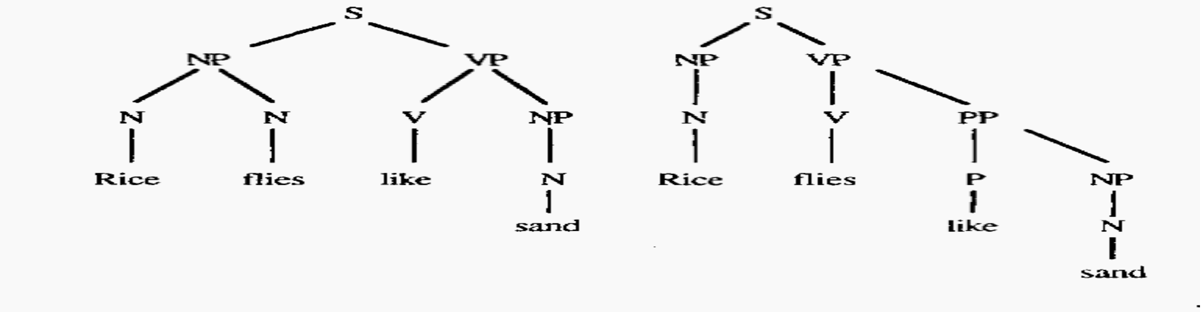


Figure 1.4 Two structural representations of *Rice flies like sand.*

• This information is often presented in a tree form
• Figure shows two different structures for the sentence "**Rice flies like sand".**
• In the first reading, the sentence is formed from a noun phrase (NP) describing a type of fly' rice flies, and a verb phrase (VP) that asserts that these flies like sand.
• In the second structure, the sentence is formed from a noun phrase describing a type of substance, rice, and a verb phrase stating that this substance flies like sand (say, if you throw it).

- The two structures also give further details on the structure of the noun phrase and verb phrase and identify the part of speech for each word.
- In particular, the word "like" is a verb (V) in the first reading and a preposition (P) in the second.

The Logical Form

- The structure of a sentence doesn't reflect its meaning
- For example, the NP "the catch" can have different meanings depending on whether the speaker is talking about a baseball game or a fishing expedition.
- Both these interpretations have the same syntactic structure, and the different meanings arise from an ambiguity concerning the sense of the word "catch".
- Once the correct sense is identified, say the fishing sense, there still is a problem in determining what fish are being referred to.
- The intended meaning of a sentence depends on the situation in which the sentence is produced.

The Final Meaning Representation

The final representation needed is a general knowledge representation (KR), which the system uses to represent and reason about its application domain. This is the language in which all the specific knowledge based on the application is represented.

In a question-answering application, a question might map to a database query, in a story-understanding application, a sentence might map into a set of expressions that represent the situation that the sentence describes.

[6]The Organization of Natural Language Understanding Systems:

- organized around the three levels of representation just discussed:
  - Syntactic structure,
  - Logical form
  - Final meaning representation.
- The below Figure shows the organization
- There are interpretation processes that map from one representation to the other.
- For instance, the process that maps a sentence to its syntactic structure and logical form is called the parser.
- It uses knowledge about word and word meanings (the lexicon) and a set of rules defining the legal structures (the grammar) in order to assign a syntactic structure and a logical form to an input sentence.
- An alternative organization could perform syntactic processing first and then perform semantic interpretation on the resulting structures.
- Since every proposed interpretation must simultaneously be syntactically and semantically well formed.

consider the following two sentences:

- *10. Visiting relatives can be trying.*
- *11. Visiting museums can be trying.*

These two sentences have identical syntactic structure, so both are syntactically ambiguous.

- In sentence 10, the subject might be relatives who are visiting you or the event of you visiting relatives.
- Both of these alternatives are semantically valid, and you would need to determine the appropriate sense by using the con -textual mechanism.
- Sentence 11 has only one possible semantic interpretation, since museums are not objects that can visit other people; rather they must be visited.

The Flow of Information

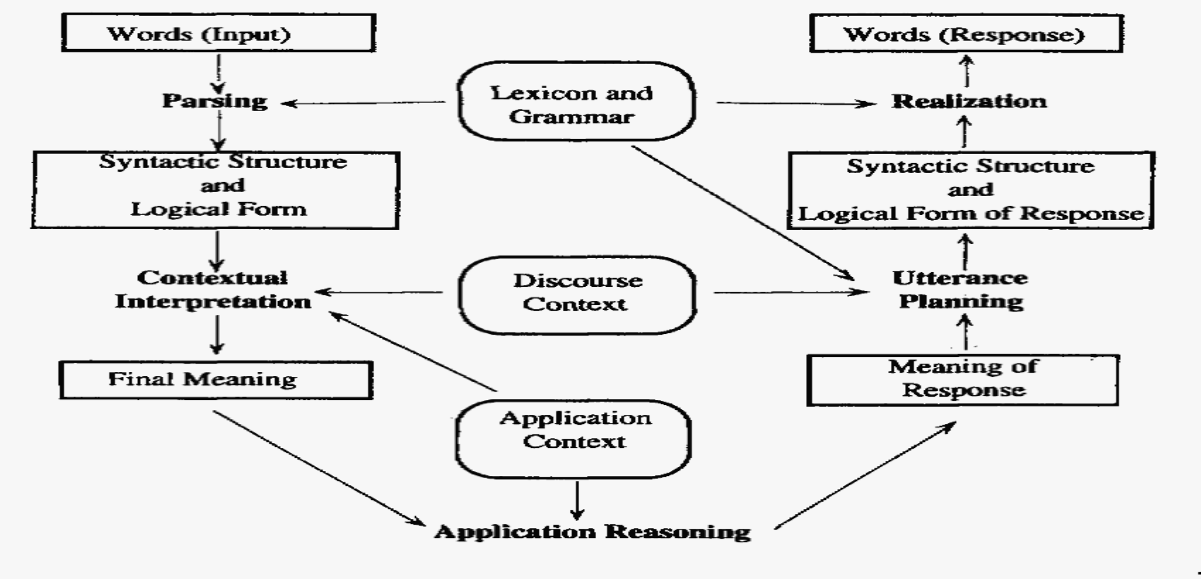


Figure 1.5 The flow of information

The Flow of Information

- The process that transforms the syntactic structure and logical form into a final meaning representation is called contextual processing.
- This process includes issues such as identifying the objects referred to by noun phrases and pronouns, the analysis of the temporal aspects of the new information conveyed by the sentence, the identification of the speaker's intention (for example, whether "**Can you lift that rock**" is a yes/no question or a request), as well as all the inferential processing required to interpret the sentence appropriately within the application domain.
- It uses knowledge of the discourse context (determined by the sentences that preceded the current one) and knowledge of the application to produce a final representation.

[7]Linguistic Background: An outline of English Syntax:

- Words
- The Elements of Simple Noun Phrases
- Verb Phrases and Simple Sentences
- Noun Phrases
- Adjective Phrases
- Adverbial Phrases