NLP: Introduction

A language is not just words. It's a culture, a tradition, a unification of a community, a whole history that creates what a community is. It's all embodied in a language.

Noam Chomsky

Imagine a hypothetical person, John Doe. He's the CTO of a fast-growing technology startup. On a busy day, John wakes up and has this conversation with his digital assistant:

John: "How is the weather today?"

Digital assistant: "It is 37 degrees centigrade outside with no rain today."

John: "What does my schedule look like?"

Digital assistant: "You have a strategy meeting at 4 p.m. and an all-hands at 5:30 p.m. Based on today's traffic situation, it is recommended you leave for the office by 8:15 a.m."

While he's getting dressed, John probes the assistant on his fashion choices:

John: "What should I wear today?"

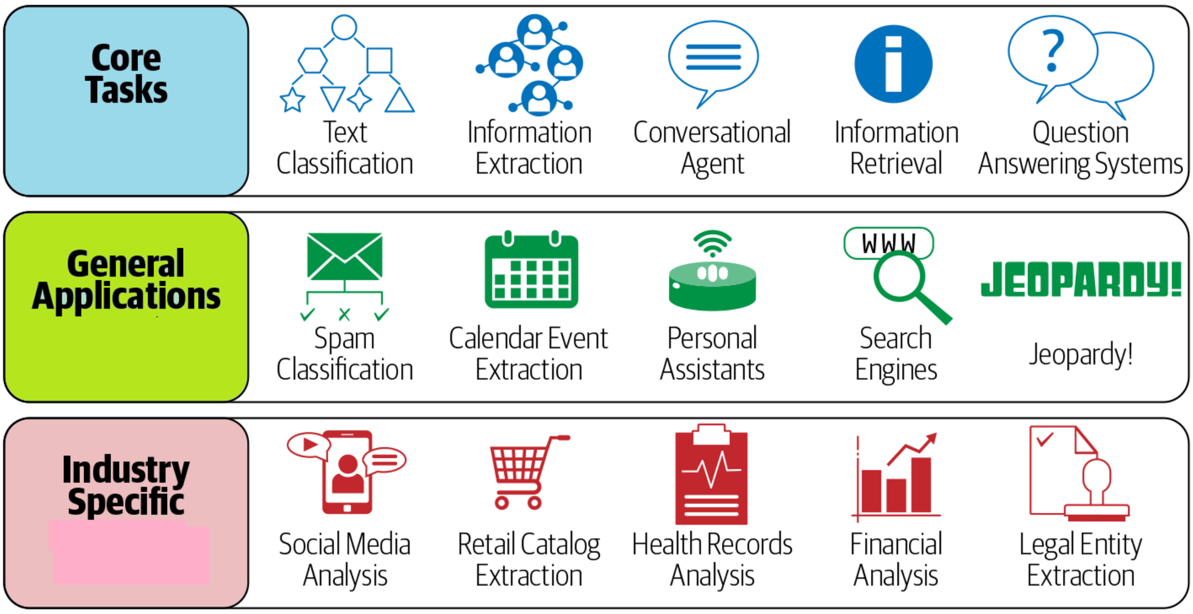
Digital assistant: "White seems like a good choice."

You might have used smart assistants such as Amazon Alexa, Google Home, or Apple Siri to do similar things.

We talk to these assistants not in a programming language, but in our natural language—the language we all communicate in.

This natural language has been the primary medium of communication between humans since time immemorial. But computers can only process data in binary, i.e., 0s and 1s. While we can represent language data in binary, how do we make machines understand the language?

This is where natural language processing (NLP) comes in. It is an area of computer science that deals with methods to analyze, model, and understand human language. Every intelligent application involving human language has some NLP behind it.



NLP in the Real World

NLP is an important component in a wide range of software applications that we use in our daily lives.

Core applications:

- Email platforms, such as Gmail, Outlook, etc., use NLP extensively to provide a range of product features, such as spam classification, priority inbox, calendar event extraction, auto-complete, etc.
- Voice-based assistants, such as Apple Siri, Google Assistant, Microsoft Cortana, and Amazon Alexa rely on a range of NLP techniques to interact with the user, understand user commands, and respond accordingly.
- Modern search engines, such as Google and Bing, which are the cornerstone of today’s internet, use NLP heavily for various subtasks, such as query understanding, query expansion, question answering, information retrieval, and ranking and grouping of the results.
- Machine translation services, such as Google Translate, Bing Microsoft Translator, and Amazon Translate are increasingly used in today’s world

to solve a wide range of scenarios and business use cases. These services are direct applications of NLP.

Other applications:

- Organizations across verticals analyze their social media feeds to build a better and deeper understanding of the voice of their customers.
- NLP is widely used to solve diverse sets of use cases on e-commerce platforms like Amazon. These vary from extracting relevant information from product descriptions to understanding user reviews.
- Advances in NLP are being applied to solve use cases in domains such as healthcare, finance, and law.
- Companies such as Arria are working to use NLP techniques to automatically generate reports for various domains, from weather forecasting to financial services.
- NLP forms the backbone of spelling- and grammar-correction tools, such as Grammarly and spell check in Microsoft Word and Google Docs.
- *Jeopardy!* is a popular quiz show on TV. In the show, contestants are presented with clues in the form of answers, and the contestants must phrase their responses in the form of questions. IBM built the Watson AI to compete with the show's top players. Watson won the first prize with a million dollars, more than the world champions. Watson AI was built using NLP techniques and is one of the examples of NLP bots winning a world competition.
- NLP is used in a range of learning and assessment tools and technologies, such as automated scoring in exams like the Graduate Record Examination (GRE), plagiarism detection (e.g., Turnitin), intelligent tutoring systems, and language learning apps like Duolingo.
- NLP is used to build large knowledge bases, such as the Google Knowledge Graph, which are useful in a range of applications like search and question answering.

NLP Tasks

There is a collection of fundamental tasks that appear frequently across various NLP projects. Owing to their repetitive and fundamental nature, these tasks have been studied extensively.

Language modeling

This is the task of predicting what the next word in a sentence will be based on the history of previous words. The goal of this task is to learn the probability of a sequence of words appearing in a given language. Language modeling is useful for building solutions for a wide variety of problems, such as speech recognition, optical character recognition, handwriting recognition, machine translation, and spelling correction.

Text classification

This is the task of bucketing the text into a known set of categories based on its content. Text classification is by far the most popular task in NLP and is used in a variety of tools, from email spam identification to sentiment analysis.

Information extraction

As the name indicates, this is the task of extracting relevant information from text, such as calendar events from emails or the names of people mentioned in a social media post.

Information retrieval

This is the task of finding documents relevant to a user query from a large collection. Applications like Google Search are well-known use cases of information retrieval.

Conversational agent

This is the task of building dialogue systems that can converse in human languages. Alexa, Siri, etc., are some common applications of this task.

Text summarization

This task aims to create short summaries of longer documents while retaining the core content and preserving the overall meaning of the text.

Question answering

This is the task of building a system that can automatically answer questions posed in natural language.

Machine translation

This is the task of converting a piece of text from one language to another. Tools like Google Translate are common applications of this task.

Topic modeling

This is the task of uncovering the topical structure of a large collection of documents. Topic modeling is a common text-mining tool and is used in a wide range of domains, from literature to bioinformatics.

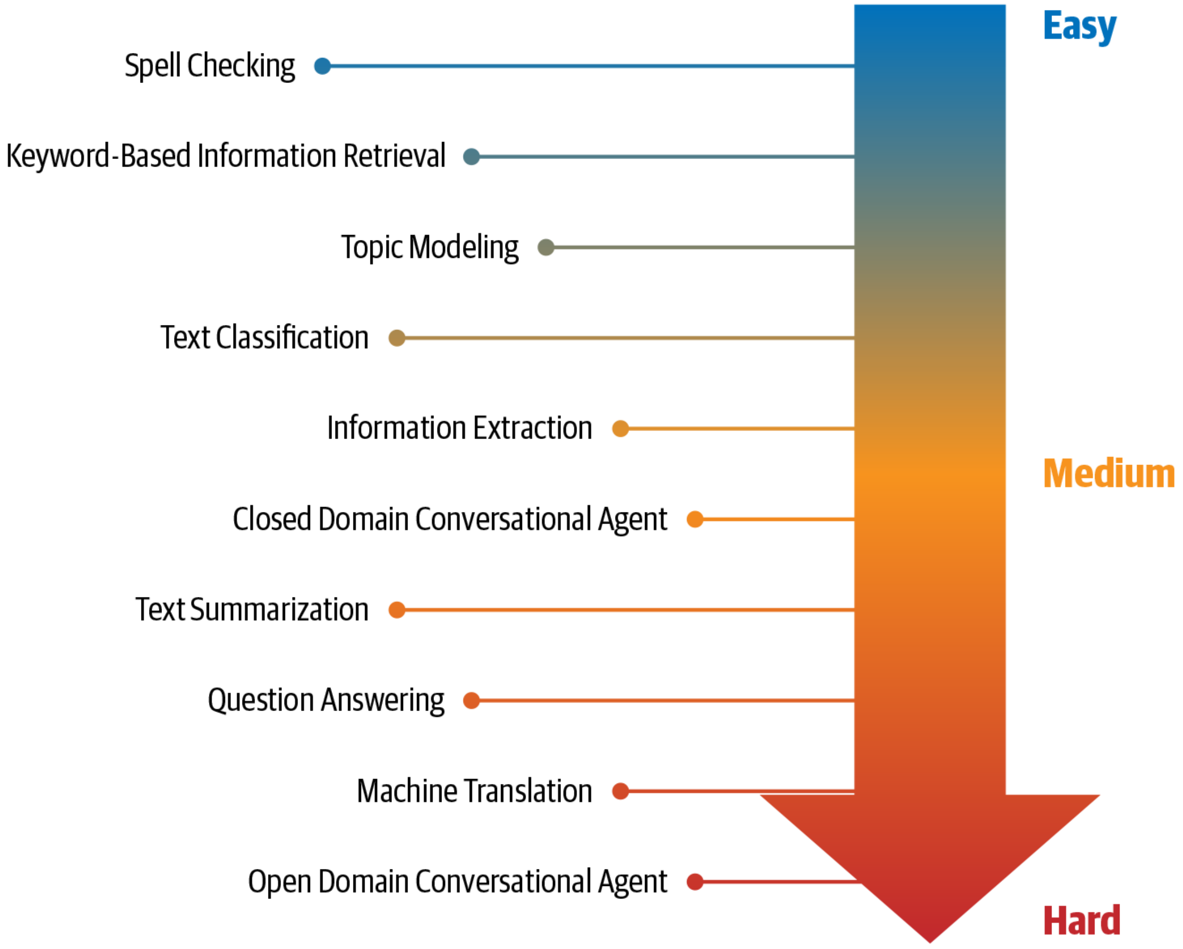


Fig above shows a depiction of these tasks based on their relative difficulty in terms of developing comprehensive solutions.

What Is Language?

Language is a structured system of communication that involves complex combinations of its constituent components, such as characters, words, sentences, etc. Linguistics is the systematic study of language. In order to study NLP, it is important to understand some concepts from linguistics about how language is structured.

We can think of human language as composed of four major building blocks: **phonemes, morphemes and lexemes, syntax, and context.**

NLP applications need knowledge of different levels of these building blocks, starting from the basic sounds of language (phonemes) to texts with some meaningful expressions (context).

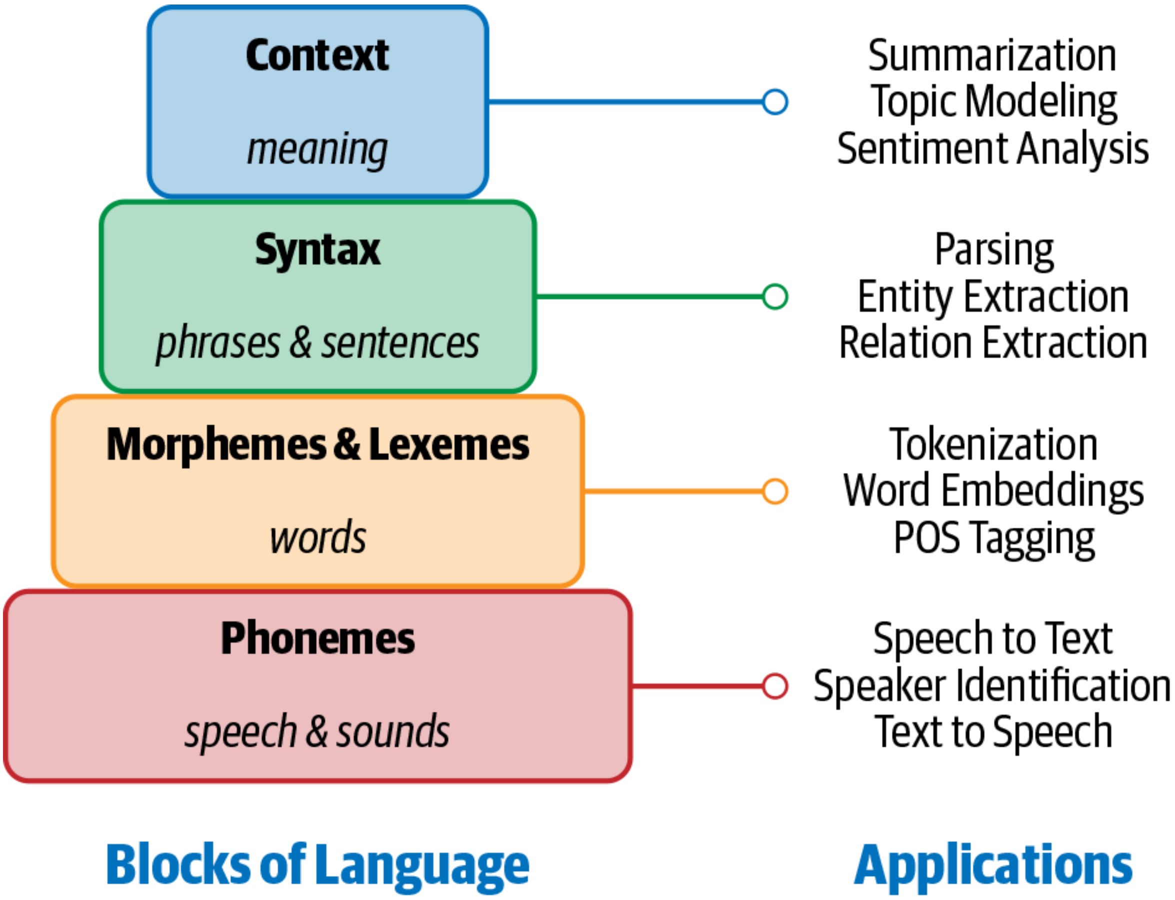


Fig: Building blocks of language and their applications

Building Blocks of Language

Let's first introduce what these blocks of language are to give context for the challenges involved in NLP.

Phonemes

Phonemes are the smallest units of sound in a language. They may not have any meaning by themselves but can induce meanings when uttered in combination with other phonemes.

For example, Standard English has 44 phonemes, which are either single letters or a combination of letters.

Figure below shows these phonemes along with sample words. Phonemes are particularly important in applications involving speech understanding, such as speech recognition, speech-to-text transcription, and text-to-speech conversion.

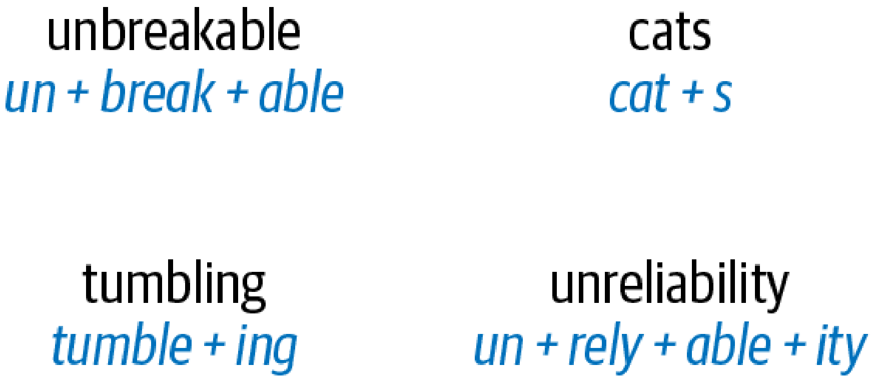
Consonant phonemes, with sample words		Vowel phonemes, with sample words	
1. /b/ - bat	13. /s/ - sun	1. /a/ - ant	13. /oi/ - coin
2. /k/ - cat	14. /t/ - tap	2. /e/ - egg	14. /ar/ - farm
3. /d/ - dog	15. /v/ - van	3. /i/ - in	15. /or/ - for
4. /f/ - fan	16. /w/ - wig	4. /o/ - on	16. /ur/ - hurt
5. /g/ - go	17. /y/ - yes	5. /u/ - up	17. /air/ - fair
6. /h/ - hen	18. /z/ - zip	6. /ai/ - rain	18. /ear/ - dear
7. /j/ - jet	19. /sh/ - shop	7. /ee/ - feet	19. /ure/[4] - sure
8. /l/ - leg	20. /ch/ - chip	8. /igh/ - night	20. /ə/ - corner (the 'schwa' - an unstressed vowel sound which is close to /u/)
9. /m/ - map	21. /th/ - thin	9. /oa/ - boat	
10. /n/ - net	22. /**th**/ - then	10. /**oo**/ - boot	
11. /p/ - pen	23. /ng/ - ring	11. /oo/ - look	
12. /r/ - rat	24. /zh/[3] - vision	12. /ow/ - cow	

Morphemes and lexemes

A morpheme is the smallest unit of language that has a meaning. It is formed by a combination of phonemes.

Not all morphemes are words, but all prefixes and suffixes are morphemes. For example, in the word "multimedia," "multi-" is not a word but a prefix that changes the meaning when put together with "media." "Multi-" is a morpheme.

Figure below illustrates some words and their morphemes. For words like "cats" and "unbreakable," their morphemes are just constituents of the full word, whereas for words like "tumbling" and "unreliability," there is some variation when breaking the words down into their morphemes.



Lexemes are the structural variations of morphemes related to one another by meaning.

For example, “run” and “running” belong to the same lexeme form. Morphological analysis, which analyses the structure of words by studying its morphemes and lexemes, is a foundational block for many NLP tasks, such as tokenization, stemming, learning word embeddings, and part-of-speech tagging.

Syntax

Syntax is a set of rules to construct grammatically correct sentences out of words and phrases in a language.

Syntactic structure in linguistics is represented in many different ways. A common approach to representing sentences is a parse tree. Figure below shows an example parse tree for two English sentences.

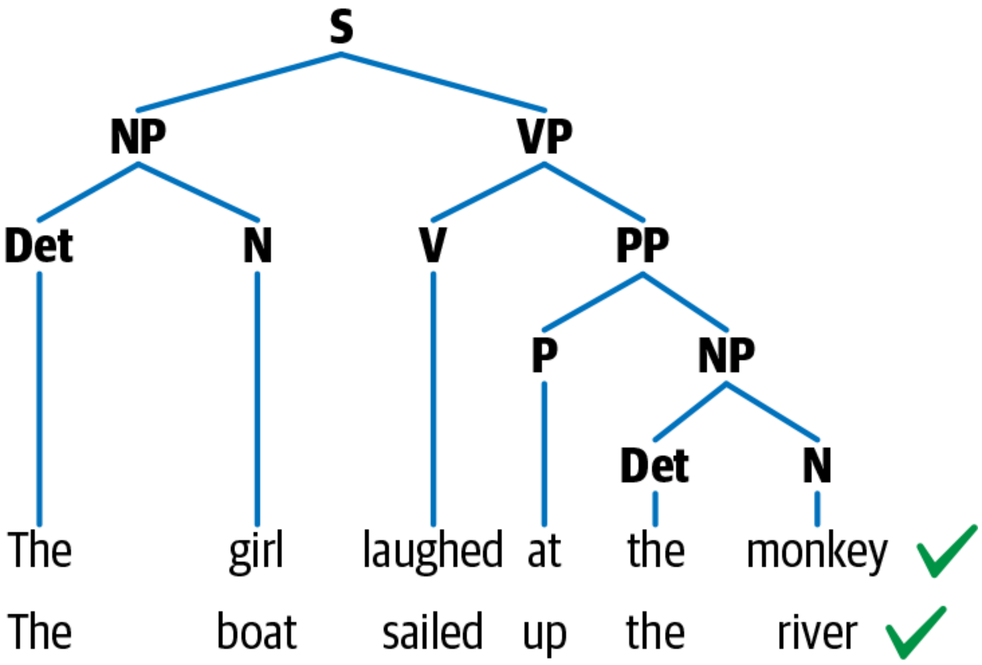

S
NP
VP
Det
N
V
PP
P
NP
Det
N
The girl laughed at the monkey ✓
The boat sailed up the river ✓

Figure: Syntactic structure of two syntactically similar sentences

This has a hierarchical structure of language, with words at the lowest level, followed by part-of-speech tags, followed by phrases, and ending with a sentence at the highest level.

In Figure above, both sentences have a similar structure and hence a similar syntactic parse tree.

In this representation, N stands for noun, V for verb, and P for preposition. Noun phrase is denoted by NP and verb phrase by VP.

The two noun phrases are “The girl” and “The boat,” while the two verb phrases are “laughed at the monkey” and “sailed up the river.”

The syntactic structure is guided by a set of grammar rules for the language (e.g., the sentence comprises an NP and a VP), and this in turn guides some of the fundamental tasks of language processing, such as parsing.

Parsing is the NLP task of constructing such trees automatically.

Entity extraction and relation extraction are some of the NLP tasks that build on this knowledge of parsing, Note that the parse structure described above is specific to English. The syntax of one language can be very different from that of another language, and the language-processing approaches needed for that language will change accordingly.

Context

Context is how various parts in a language come together to convey a particular meaning.

Context includes long-term references, world knowledge, and common sense along with the literal meaning of words and phrases.

The meaning of a sentence can change based on the context, as words and phrases can sometimes have multiple meanings.

Generally, context is composed from semantics and pragmatics.

Semantics is the direct meaning of the words and sentences without external context.

Pragmatics adds world knowledge and external context of the conversation to enable us to infer implied meaning.

Complex NLP tasks such as sarcasm detection, summarization, and topic modeling are some of tasks that use context heavily.

Why Is NLP Challenging?

What makes NLP a challenging problem domain? The ambiguity and creativity of human language are just two of the characteristics that make NLP a demanding area to work in.

Ambiguity

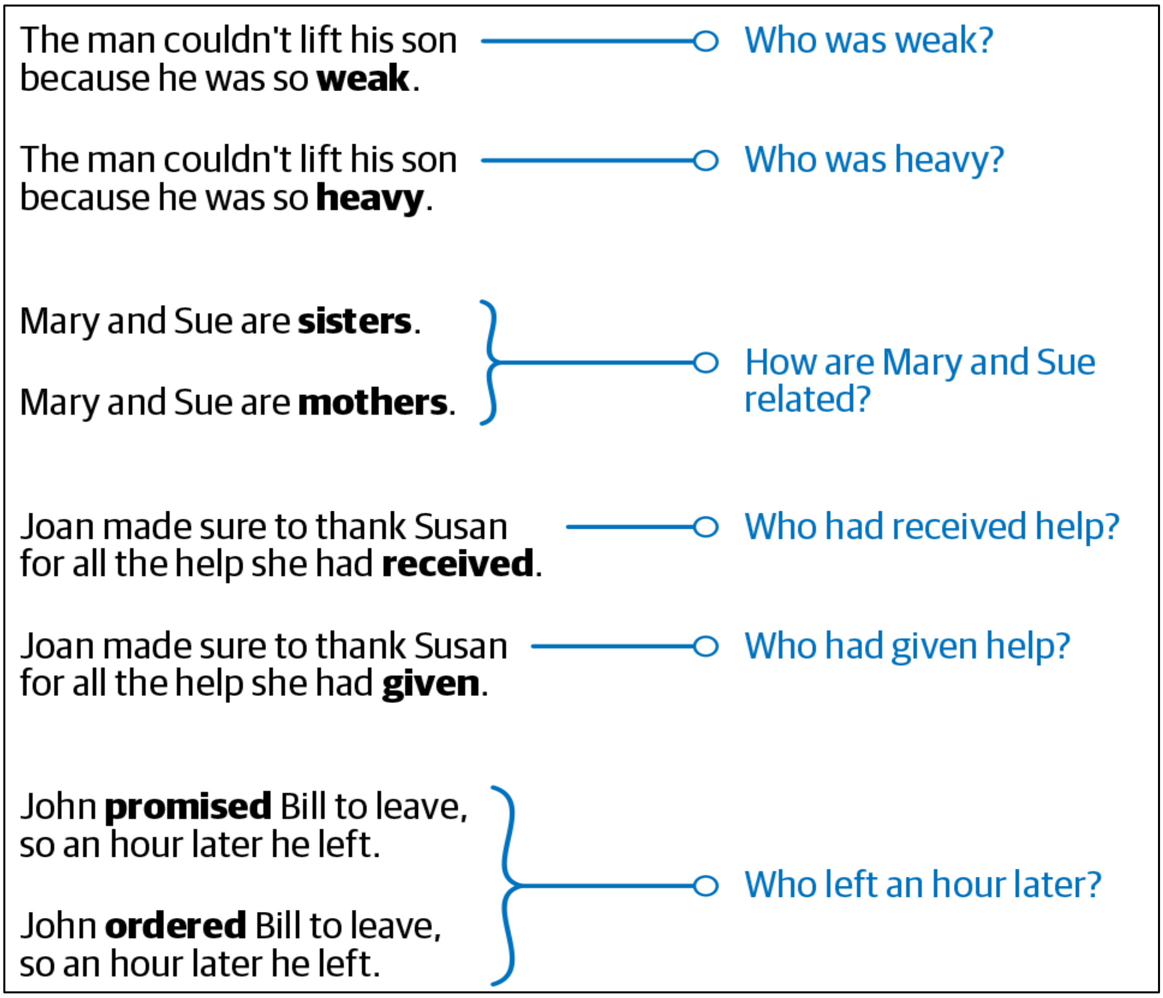
Ambiguity means uncertainty of meaning. Most human languages are inherently ambiguous.

Consider the following sentence: “I made her duck.” This sentence has multiple meanings. The first one is: I cooked a duck for her. The second meaning is: I made her bend down to avoid an object.

Here, the ambiguity comes from the use of the word “made.” Which of the two meanings applies depends on the context in which the sentence appears.

If the sentence appears in a story about a mother and a child, then the first meaning will probably apply. But if the sentence appears in a book about sports, then the second meaning will likely apply. The example we saw is a direct sentence.

When it comes to figurative language—i.e., idioms—the ambiguity only increases. For example, “He is as good as John Doe.” Try to answer, “How good is he?” The answer depends on how good John Doe is. Figure below shows some examples illustrating ambiguity in language.



The man couldn't lift his son because he was so **weak**.

Who was weak?

The man couldn't lift his son because he was so **heavy**.

Who was heavy?

Mary and Sue are **sisters**.

Mary and Sue are **mothers**.

How are Mary and Sue related?

Joan made sure to thank Susan for all the help she had **received**.

Who had received help?

Joan made sure to thank Susan for all the help she had **given**.

Who had given help?

John **promised** Bill to leave, so an hour later he left.

John **ordered** Bill to leave, so an hour later he left.

Who left an hour later?

The examples come from the Winograd Schema Challenge, named after Professor Terry Winograd of Stanford University.

This schema has pairs of sentences that differ by only a few words, but the meaning of the sentences is often flipped because of this minor change.

These examples are easily disambiguated by a human but are not solvable using most NLP techniques.

Consider the pairs of sentences in the figure and the questions associated with them. With some thought, how the answer changes should be apparent based on a single word variation.

- **Components of NLP**

There are two components of NLP as given −

- **Natural Language Understanding (NLU)**

Understanding involves the following tasks −

- Mapping the given input in natural language into useful representations.
- Analyzing different aspects of the language.

- **Natural Language Generation (NLG)**
    - It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation.

It involves −

- **Text planning** − It includes retrieving the relevant content from knowledge base.
- **Sentence planning** − It includes choosing required words, forming meaningful phrases, setting tone of the sentence.
- **Text Realization** − It is mapping sentence plan into sentence structure.

The NLU is harder than NLG.

Difficulties in NLU

Natural Language has an extremely rich form and structure.

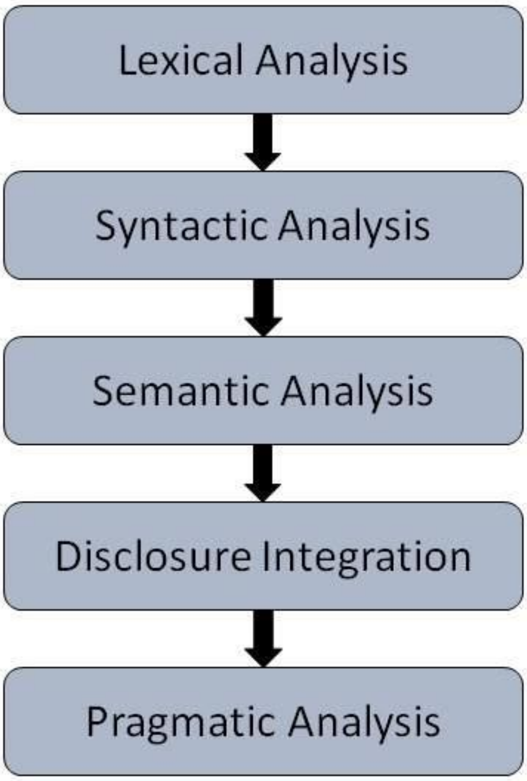
It is very ambiguous. There can be different levels of ambiguity −

- **Lexical ambiguity** − It is at very primitive level such as word-level.
- For example, treating the word "board" as noun or verb?
- **Syntax Level ambiguity** − A sentence can be parsed in different ways.
- For example, "He lifted the beetle with red cap." − Did he use cap to lift the beetle or he lifted a beetle that had red cap?
- **Referential ambiguity** − Referring to something using pronouns. For example, Rima went to Gauri. She said, "I am tired." − Exactly who is tired?
- One input can mean different meanings.
- Many inputs can mean the same thing.

NLP Terminology

- **Phonology** – It is study of organizing sound systematically.
- **Morphology** – It is a study of construction of words from primitive meaningful units.
- **Morpheme** – It is primitive unit of meaning in a language.
- **Syntax** – It refers to arranging words to make a sentence. It also involves determining the structural role of words in the sentence and in phrases.
- **Semantics** – It is concerned with the meaning of words and how to combine words into meaningful phrases and sentences.
- **Pragmatics** – It deals with using and understanding sentences in different situations and how the interpretation of the sentence is affected.
- **Discourse** – It deals with how the immediately preceding sentence can affect the interpretation of the next sentence.
- **World Knowledge** – It includes the general knowledge about the world.

Steps/Stages in NLP



There are general five steps −

- **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 txt into paragraphs, sentences, and words.
- **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 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".
- **Discourse Integration** − The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.
- **Pragmatic Analysis** − 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.