Natural Language Processing

Contents

- Introduction
- Definition
- Components of NLP
- Overview of Linguistics
- Grammars and Languages
- What is NLP used for?
- History of NLP

Introduction

- Natural Language: How humans communicate with each other. Computers should replicate same thing.
- Developing programs to understand natural language is important in AI.
- Natural form of communication with systems is essential for user acceptance.
- One of the critical tests for intelligent behaviour is the ability to communicate effectively.

Definition

- NLP involves machines and robots to understand and process the language that humans speak.

- A typical interaction between humans and machines using Natural Language Processing could go as follows:

1. 1. A human talks to the machine
2. 2. The machine captures the audio
3. 3. Audio to text conversion takes place
4. 4. Processing of the text's data
5. 5. Data to audio conversion takes place
6. 6. The machine responds to the human by playing the audio file

What is NLP used for?

Natural Language Processing is the driving force behind the following common applications:

- Language translation applications such as Google Translate
- Word Processors such as Microsoft Word and Grammarly that employ NLP to check grammatical accuracy of texts.
- Interactive Voice Response (IVR) applications used in call centers to respond to certain users' requests.
- Personal assistant applications such as OK Google, Siri, Cortana, and Alexa.
- Chatbots etc.

History of NLP

- Here, are important events in the history of Natural Language Processing:
- **1950-** NLP started when Alan Turing published an article called "Machine and Intelligence."
- **1950-** Attempts to automate translation between Russian and English
- **1960-** The work of Chomsky and others on formal language theory and generative syntax
- **1990-** Probabilistic and data-driven models had become quite standard
- **2000-** A Large amount of spoken and textual data become available

How does NLP work?

- Let's understand how humans use language.
- Every day, we say thousands of words that other people interpret to do countless things.
- We, consider it as a simple communication, but we all know that words run much deeper than that.
- There’s always some context that we derive from what we say and how we say it. NLP never focuses on voice modulation; it does draw on contextual patterns.

- **Example:**

Man is to woman as king is to _____?

Meaning (king) – meaning (man) + meaning (woman)=?

The answer is- queen

Here, we can easily co-relate because man is male gender and woman is female gender. In the same way, the king is masculine gender, and its female gender is queen.

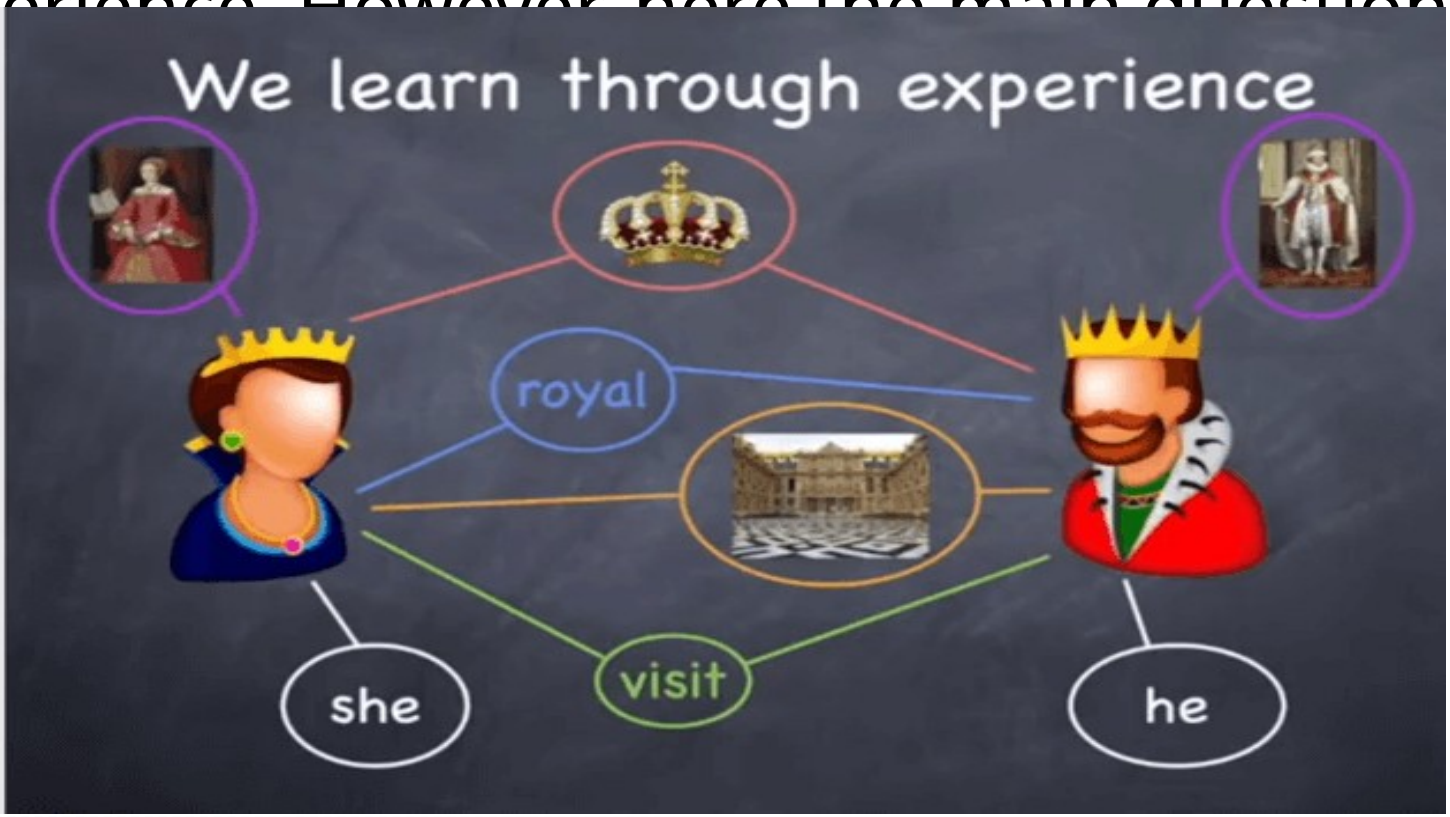
- **Example:**

Is King to kings as the queen is to_____?

The answer is- queens

Here, we can see two words kings and kings where one is singular and other is plural. Therefore, when the word queen comes, it automatically co-relates with queens again singular plural.

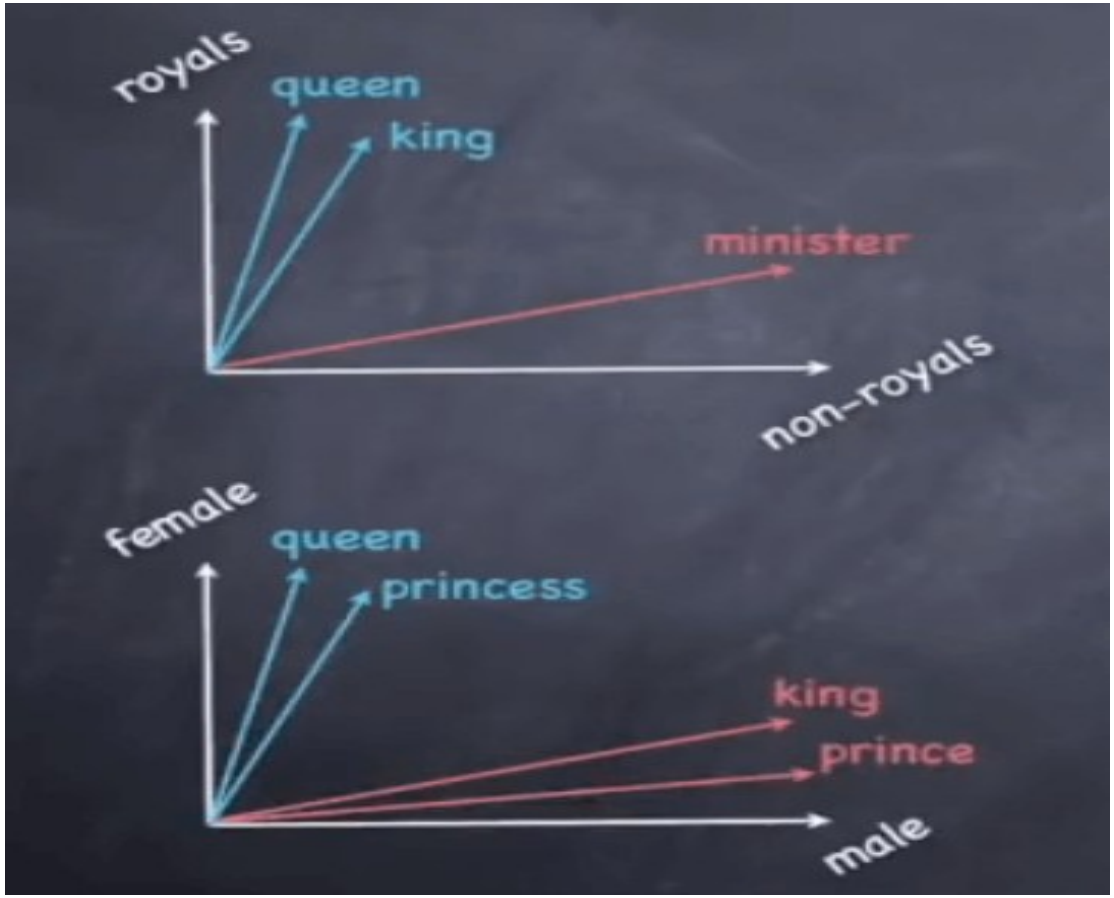
- Here, the biggest question is that how do we know what words mean? Let's, say who will call it queen?
- The answer is we learn this thinks through experience. However here the main question is tha



- We need to provide enough data for Machines to learn through experience. We can feed details like:
  - Her Majesty the Queen.
  - The Queen's speech during the State visit
  - The crown of Queen Elizabeth
  - The Queens's Mother
  - The queen is generous.

- With above examples the machine understands the entity Queen.

- The machine creates word vectors as below. A word vector is built using surrounding words.



- The machine creates these vectors
  - As it learns from multiple datasets
  - Use Machine learning (e.g., Deep Learning algorithms)
  - A word vector is built using surrounding words.

- Here is the formula:

  Meaning (king) – meaning (man) + meaning (woman)=?

- This amounts to performing simple algebraic operations on word vectors:

  Vector (king) – vector (man) + vector (woman)= vector(?)

- To which the machine answers queen.

Components of NLP

1. **Natural Language Understanding**
   - Taking some spoken/typed sentence and working out

     what it means.
   - Challenges: ambiguity
     - Lexical ambiguity (word level e.g. Fly, bear, can have different meanings)
     - Syntactical ambiguity (e.g. Old men and women were taken to safe place.)
     - Semantic ambiguity (different meanings eg. The car hit the pole while it was moving.)
     - Referential ambiguity (e.g. Sita went to Gita. She said, "I am hungry.")

2. Natural Language Generation

- Taking some formal representation of what you want to say and working out a way to express it in a natural (human) language (e.g., English)

- Involves following steps:
  - Text planning
  - Sentence planning
  - Text realization

Overview of Linguistics

- Scientific study of language.
- In a natural language the sentence is the basic language element.
- Sentence made up of words which express a complete thought.
- To express a complete thought a sentence must have a subject and a predicate.
- Subject is what sentence is about
- Predicate says something about subject.

- Sentences classified by structure and usage.
- Simple sentences: one independent clause comprised of subject and predicate.
- Compound sentences: two or more independent clauses connected by conjunction or semicolon.
- Complex sentences: independent clause and one or more dependent clauses.
- Sentences used to assert, query and describe.
- A word functions in a sentence as a part of speech.
- Parts of speech for English language: noun, pronoun, verbs, adjectives, adverbs, prepositions, conjunctions, interjections.

Levels of Knowledge used in language understanding

- **Phonological:** Knowledge that relates sounds to words. Essential for speech based systems as they deal with how words are related to the sounds that realize them.

- **Morphological:** Concerns how words are constructed from morphemes. A morpheme is smallest unit of meaning.

  Ex. Construction of friendly from root *friend* and suffix *ly.*

- **Syntactic:** Relates to how words are put together or structured to form grammatically correct sentences in the language.

- **Semantic:** Concerned with meaning of words and phrases and how they combine to form sentence meanings.

- **Pragmatic:** High level knowledge which relates to the use of sentences in different contexts and how contexts affects meaning of the sentences.

- **World:** Relates to knowledge a user must have in order to understand and carry on a conversation.

Grammars and Languages

- A language L can be considered as set of strings of finite or infinite length.
- String constructed by concatenating basic atomic elements called symbols.
- Finite set v of symbols called alphabet or vocabulary.
- Among all possible strings generated from v are sentences.
- Well formed sentences constructed using set of rules called grammar.
- Language generated by grammar G denoted by L(G).

- More formally we define a grammar G as

$G=(v_n, v_t, s, p)$ where,

v_n - set of non-terminal symbols

v_t - set of terminal symbols

s- starting symbol

p-finite set of productions

or rewrite rules

- A general production rule from p has the form

$xyz \rightarrow xwz$

Where x,y,z and w are strings from v.

Example:

Q_N ={S, NP, N, VP,V, ART}

Q_T ={boy, popsicle, frog, ate, kissed, flew, the, a}

Rewrite rules given by

P:	S	NP VP
NP	ART N	→
VP	V NP	→
N	boy|popsicle|frog	→
V	ate|kissed|flew	→
ART	the|a	→

Vertical bar indicates alternative choices. S is initial symbol.

NP stands for Noun Phrase, VP stands for verb phrase, N stands for noun, V stands for verb, ART stands for article.

With this G, sentences such as following can be generated

- The boy ate a popsicle.
- The frog kissed a boy.
- A boy ate a frog.

To generate a sentence the rules from P are applied sequentially starting with S and proceeding until all non terminal symbols are eliminated.

S → NP VP

→ ART N VP

→ the N VP

→ the boy VP

→ the boy V NP

→ the boy ate NP

→ the boy ate ART N

→ the boy ate a N

→ the boy ate a popsicle