# NLP Introduction

1. It is an unsupervised learning method. For large scale, at the end to finetune we use small set as supervised.
2. **Lexical Analysis:** Normalize and disambiguated words, words with multiple meaning, multi-word expressions [make up, take out]
3. **Ambiguity:** The way words are put into the sentence; it can have multiple meanings.
4. **Pragmatic:** These are the sentences that have very different meanings

*[You: "Can you pass me the salt?",*

*Waiter: "Yes!" stares at you without flinching.]*

1. **Architecture:** It is sequence processing: RNN, CNNs, sequence-to-sequence models (like encoder & decoder), attention, transformers.
2. **Applications:** Spell check, text classification, machine translations, information extraction, chatbots, sentiment analysis.

# Basic Text Processing

1. Used to describe text patterns; information from texts; & convert text into convenient/standard form.
2. Text Normalization
   1. Tokenization
   2. Lemmatization & Stemming
   3. Logo

      Description automatically generatedSentence Segmentation
3. There are multiple ways to write a word, so we regex are used to represent them as shown:

Ways to representing regex:

Text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

* + - We can use pipe for disjunction: yours|mine; a|b|c.
    - Optional Elements and Wildcards: **? \* + .**
      * ? optional previous character
      * \* zero or more instances of previous character
      * + one or more instances of the previous charcter
      * .any character

1. Anchors:

Text

Description automatically generated