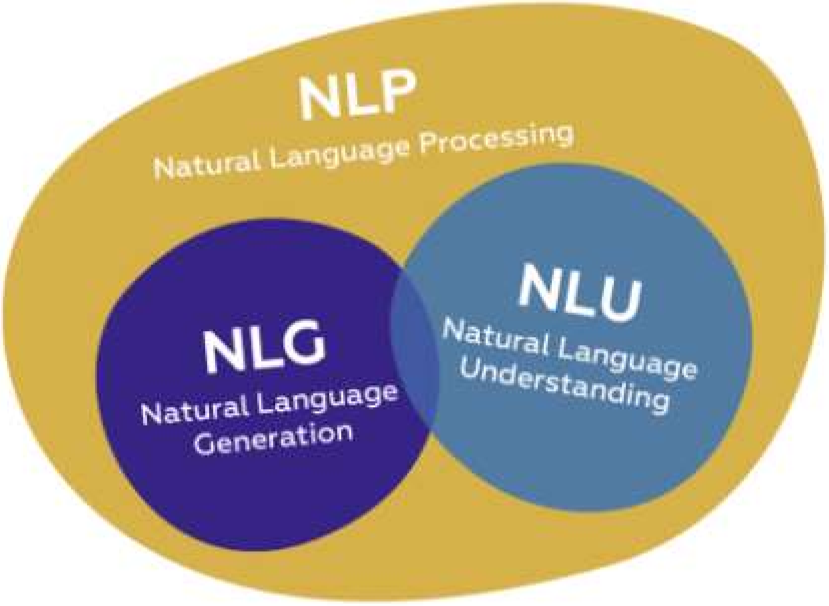
Natural Language Processing
Lab week 4 – Language modelling

This practical exercise aims to familiarize yourself with the Natural Language Generation task with the N-Gram Language Model's help.


NLP
Natural Language Processing
NLG
Natural Language Generation
NLU
Natural Language Understanding

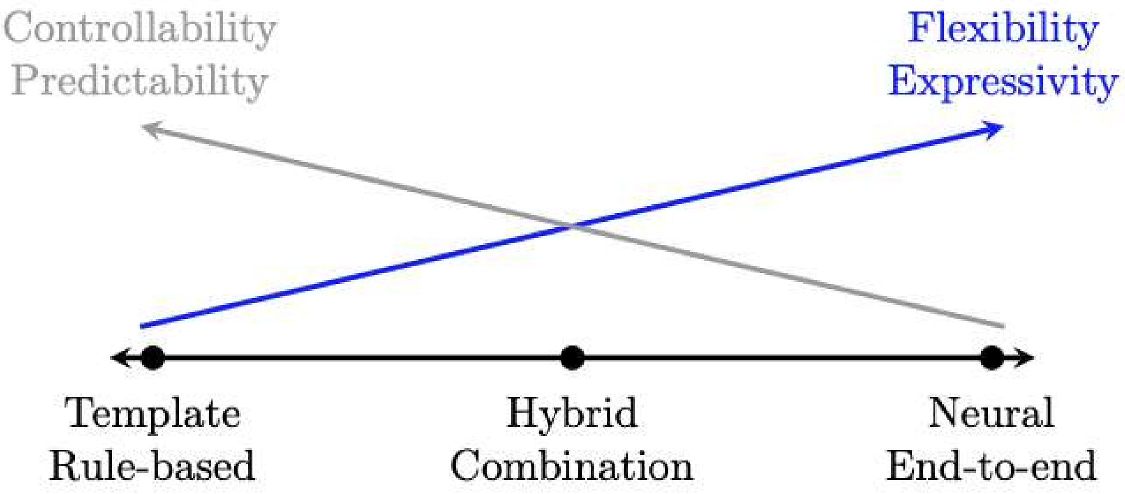
Natural Language Generation is the "process of producing meaningful phrases and sentences in the form of natural language." In its essence, it automatically generates narratives that describe, summarize, or explain input structured data in a human-like manner at the speed of thousands of pages per second.

Practical Application of NLG:

a. Analysis for business intelligence dashboards
b. Reporting on business data/ data analysis
c. IoT device status and maintenance reporting
d. Individual client financial portfolio summaries and updates
e. Personalized customer communications.

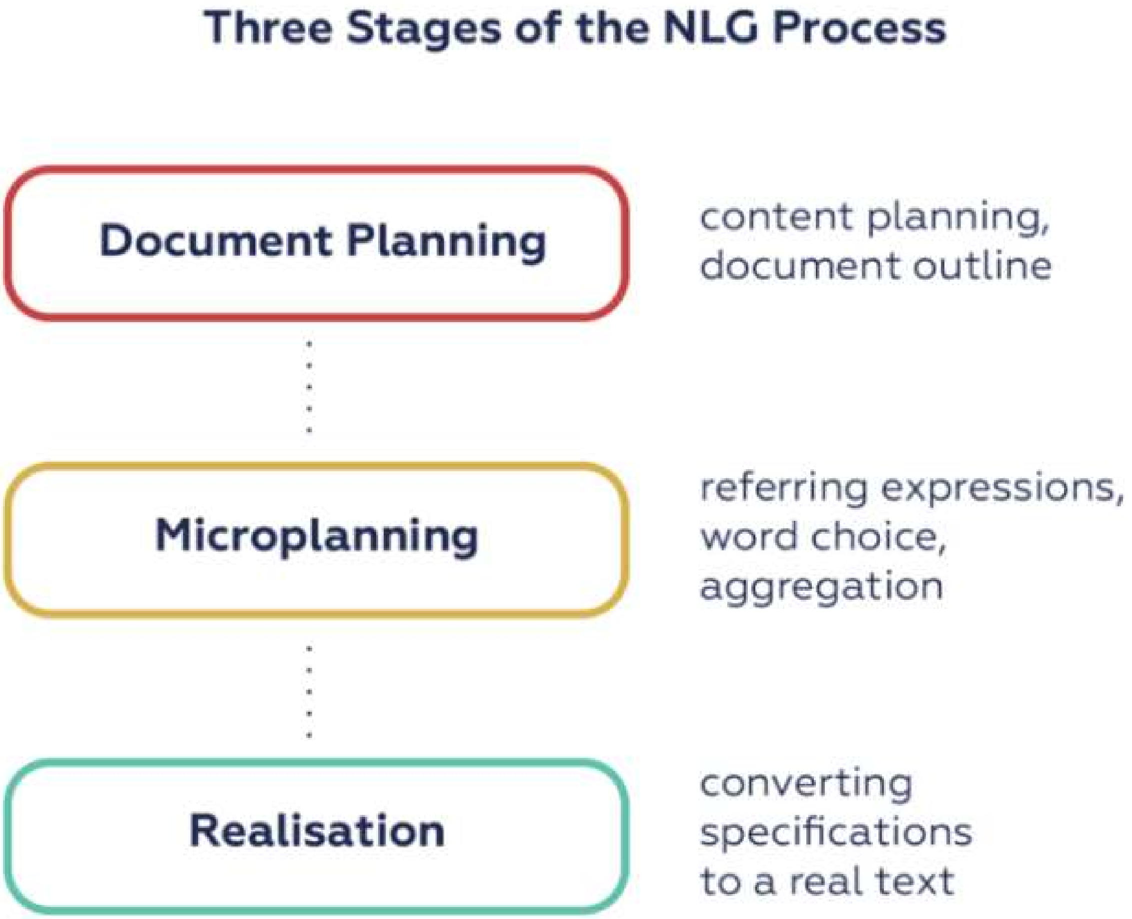
Two ends of Natural Language Generation:

There are two major approaches to language generation: using templates and dynamic creation of documents.



Stages of the NLG process:

This pipeline [ref] shows natural language generation; however, specific steps and approaches, and the models used can vary significantly with technology development.



Lab exercise:

1. **Data:**

For this lab, we will be using the underneath dataset.

a. https://github.com/praveenjoshi01/COMP9066_27375-Natural-Language-Processing/blob/main/language-never-random.txt
b. Reuters Corpus - https://www.nltk.org/book/ch02.html

2. **Document Planning:**

How should the data be formulated for ingestion into the n-gram model?

a. Removing stop words and special symbols.
b. Add special "padding" symbols to the sentence before splitting it into n-grams.
c. If required:
    a. Lemmatisation
    b. Stemming

3. **Building Language Models:**

In this step, we will be building the n-gram model to predict the next phrase or next word.

a. The n-gram model should take into account all the smaller order n-gram while building the model.
b. Add α-smoothing.

4. **Building Application:**

Write a function, which will take underneath parameters:

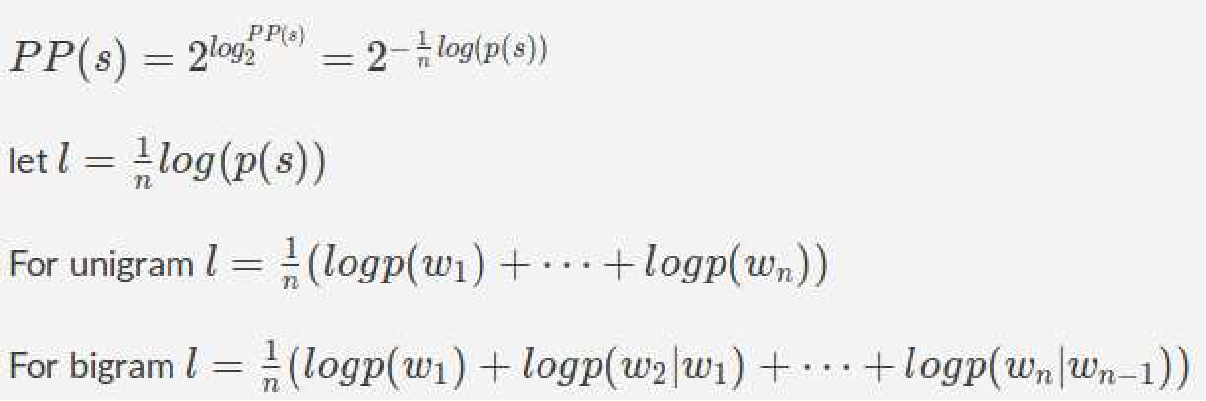
a. N-Gram model
b. Number of words to generate
c. Random seed

The function will be responsible for generating the text of given words(b) when provided with the n-gram model(a) and random seed (c).

5. **Evaluation:**

The quality of the language model should be evaluated using perplexity.

$$PP = 2^{H(P_{LM})}$$



$$PP(s) = 2^{log_2^{PP(s)}} = 2^{-\frac{1}{n}log(p(s))}$$

let $l = \frac{1}{n}log(p(s))$

For unigram $l = \frac{1}{n}(logp(w_1) + \cdots + logp(w_n))$

For bigram $l = \frac{1}{n}(logp(w_1) + logp(w_2|w_1) + \cdots + logp(w_n|w_{n-1}))$