Année Universitaire : 2025/2026 Master 2 : SII Module : TALN	Université des Sciences et de la Technologie Houari Boumediene Faculté d'Informatique Département d'Intelligence Artificielle et Sciences des Données	**TP N°3** N-Gram Language Model Part 2

I. Estimating the Next Word in a Sequence of Words

- Using an **N-gram language model**, the goal is to estimate the most probable next word in a given sequence.
- The general procedure for estimating the next word (see page 2 for the example) is based on the **interpolation method** covered in the course, as follows:

Case 1 – The input sequence contains only one word:

- The **Bigram model** is used to estimate the next word $\boldsymbol{w_n}$.
- The word $\boldsymbol{w_n}$ with the highest probability $\boldsymbol{P(w_n \mid w_{n-1})}$ is selected.
- In the example below, the best next word $\boldsymbol{w_n}$ is **"swarm"**, which has the highest probability $\boldsymbol{P(w_n \mid w_{n-1})}$, where $\boldsymbol{w_{n-1}}$ = **"particle"**.

Case 2 – The input sequence contains at least two words:

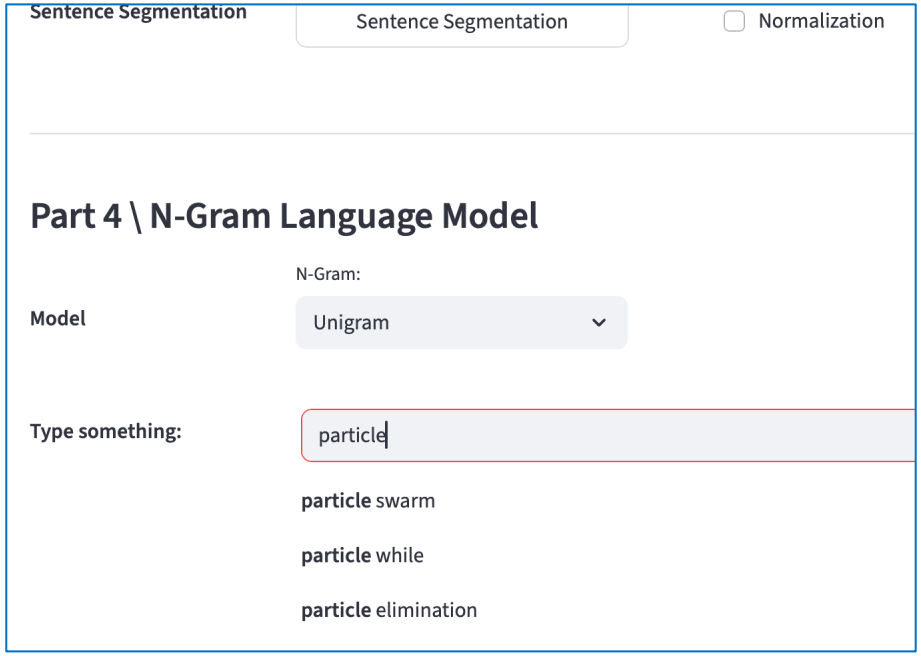
- Both the **Bigram** and **Trigram** models are used to estimate the next word $\boldsymbol{w_n}$.
- The interpolated probability is computed as:

$$\boldsymbol{P(w_n \mid w_{n-2}, w_{n-1}) = \lambda_1 P(w_n \mid w_{n-1}) + \lambda_2 P(w_n \mid w_{n-2}, w_{n-1})}$$

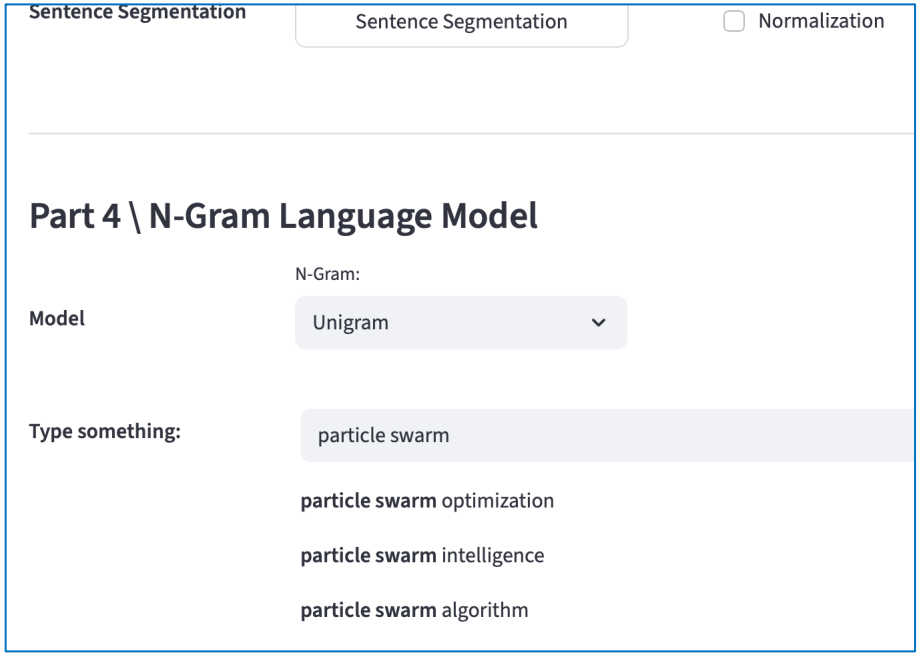
Avec :

$$\boldsymbol{\lambda_1 + \lambda_2 = 1}$$

- The word $\boldsymbol{w_n}$ with the highest interpolated probability is selected.
- In the example below, the best next word is **"optimization"**, which has the highest probability given $\boldsymbol{w_{n-2}}$ = **"particle"** and $\boldsymbol{w_{n-1}}$ = **"swarm"**


Sentence Segmentation
Sentence Segmentation
Normalization
Part 4 \ N-Gram Language Model
N-Gram:
Model
Unigram
Type something:
particle
particle swarm
particle while
particle elimination

Case 1


Sentence Segmentation
Sentence Segmentation
Normalization
Part 4 \ N-Gram Language Model
N-Gram:
Model
Unigram
Type something:
particle swarm
particle swarm optimization
particle swarm intelligence
particle swarm algorithm

Case 2

II. Estimating the Entire Sequence of Words

- Using an **N-gram language model**, the goal is to estimate the most probable entire sequence of words.
- The general procedure for estimating the entire sequence (see the example below) is based on the **Bigram model**, as follows:
  - All possible **combinations of words** provided by the user are **generated**.
  - For each combination, the **probability of the entire sequence** is computed, as follows:

$$P(w_1, w_2, \dots, w_n) = \prod_{k=1}^{n} P(w_k \mid w_{k-1})$$

  - The probability is calculated **without considering** the symbols **<s>** and **</s>**.
  - The combination of words with the **highest probability** is displayed as the **most likely correct sequence**.