INSTITUT SUPERIEUR
DE MANAGEMENT,
D'ADMINISTRATION
ET DE GENIE INFORMATIQUE
ISMAGI
RECONNU PAR L'ÉTAT

Traitement du langage naturel

Part-of-Speech (POS)

CI : Informatique (4émé Année)

2023/2024

Plan

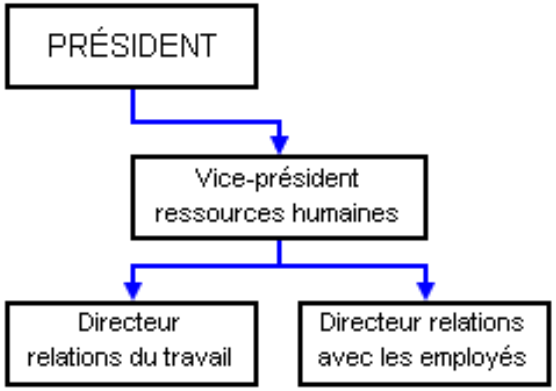
1. Introduction
2. Fondements du POS Tagging
3. Méthodes de POS Tagging
4. Évaluation du POS Tagging
5. Défis et Limitations du POS Tagging
6. Similarité sémantique du texte
7. Désambiguïsation du sens des mots

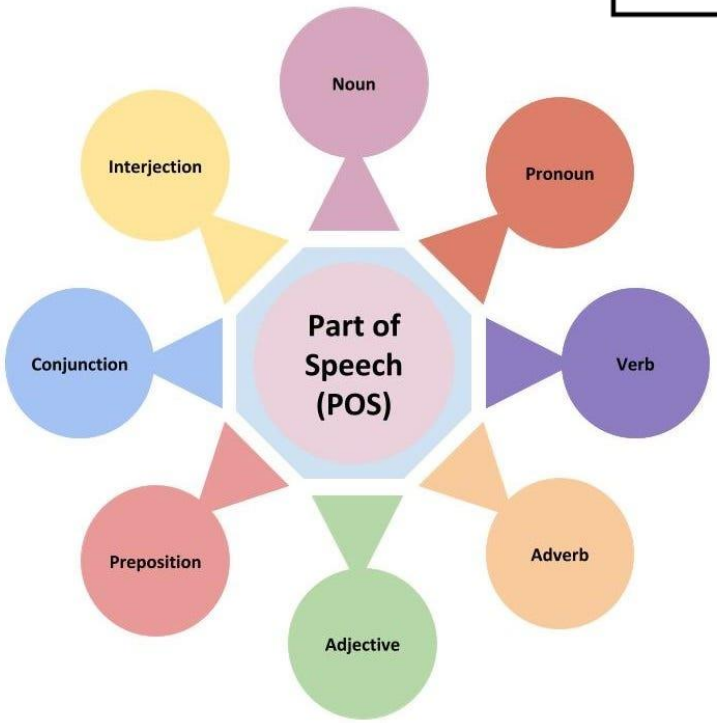
Introduction

- Les POS nous permettent de comprendre la fonction de chaque mot dans une phrase, tout comme les différentes formes et couleurs des pièces nous aident à comprendre leur place dans l'image finale.
- **POS Tagging**, une technique essentielle en traitement automatique du langage (NLP).
- Le **POS Tagging** consiste précisément à **identifier la fonction grammaticale** de chaque mot dans une phrase, en lui assignant une **étiquette** (**tag**) correspondante.

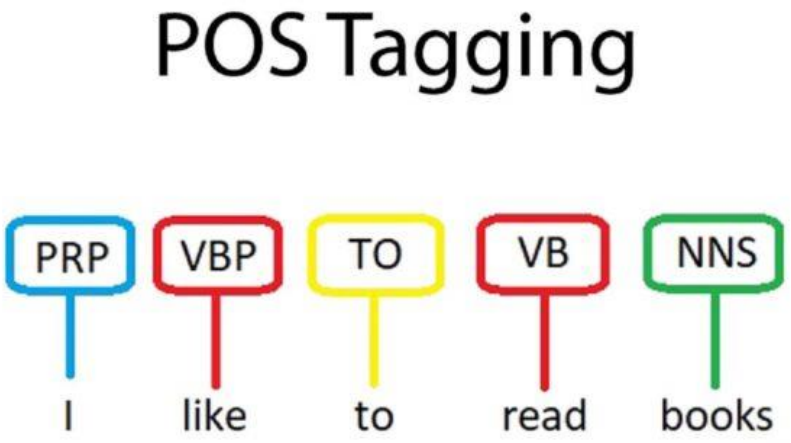
Introduction

- **Imaginez une phrase comme une entreprise.** Chaque responsable a une tâche précise : Président, Vice-…, Agent.
- Le POS Tagging permet de savoir si un mot est un **nom** (le sujet), un **verbe** (l'action), un **adjectif** (la description), etc.





Introduction



Le chat a vite mangé une souris.

Le:	article défini
chat:	nom
a mangé:	verbe
vite:	adverbe
une:	article indéfini
petite:	adjectif qualificat
souris:	nom

Fondements du POS Tagging

Alphabetical list of part-of-speech tags used in the Penn Treebank Project:

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Fondements du POS Tagging

- *Les parties du discours*
  - *LE NOM*
  - *LE DÉTERMINANT*
  - *L'ADJECTIF QUALIFICATIF*
  - *LE PRONOM*
  - *LE VERBE*
  - *L'ADVERBE*
  - *LA PRÉPOSITION*
  - *LA CONJONCTION*
  - *L'INTERJECTION*

Fondements du POS Tagging

- **LE NOM:** mot qui désigne une personne, un lieu, ou une chose (concrète ou abstraite)
- **LE DÉTERMINANT:** mot qui porte les signes du genre (féminin ou masculin) et du nombre (singulier ou pluriel) du nom qu'il accompagne et qui peut porter une précision supplémentaire
  - l'article: défini, indéfini, partitif: le, la, les, un, une, etc.
  - adjectif possessif: mon, ma, ton, ta, ses, nos, etc.
  - adjectif démonstratif: ce, cet, cette, ces
  - adjectif indéfini: aucun, chaque, autre, tel, etc.
  - adjectif interrogatif: quel, quelle, quels, quelles
  - adjectif numéral: un(e), deux, trois, quatre, etc.
  - adjectif exclamatif: quel, quelle, quels, quelles

Fondements du POS Tagging

- **L'ADJECTIF QUALIFICATIF** mot qui décrit, qualifie, une personne ou une chose
- **LE PRONOM:** mot qui remplace un nom
  - pronom personnel: je, tu, il, elle, etc.; me, te, se, lui, etc.
  - pronom possessif: le mien, la mienne, le tien, le sien, etc.
  - pronom démonstratif: celui, celle, ceux, celles
  - pronom interrogatif: lequel, laquelle, lesquels, lesquelles; qui, que, quoi
  - pronom indéfini: aucun, autre, chacun, tout, etc.
  - pronom relatif: qui, que, quoi, dont, etc.
- **LE VERBE**
- **L'ADVERBE**
- **LA PRÉPOSITION**

Fondements du POS Tagging

- Le POS Tagging repose sur des algorithmes capables d'analyser le contexte et les propriétés linguistiques des mots.
- Le POS Tagging permet de donner du sens aux phrases en révélant la fonction grammaticale de chaque mot.
- Cela ouvre la voie à de nombreuses applications NLP passionnantes :
  - **Analyse de sentiment:** Identifier le ton général d'un texte en comprenant le rôle des adjectifs et des adverbes.
  - **Extraction d'entités nommées (NER):** Repérer les entités importantes (personnes, lieux, organisations) en analysant les noms et les adjectifs qui les entourent.
  - **Traduction automatique:** Améliorer la précision de la traduction en tenant compte de la structure grammaticale de la phrase source.

Méthodes de POS Tagging

- **Méthodes Basées sur des Règles (Rule-based methods)**
  - Ces méthodes reposent sur un ensemble de **règles linguistiques** écrites à la main.
  - Les règles prennent en compte des facteurs comme la morphologie du mot (préfixes, suffixes), sa position dans la phrase, et le POS des mots voisins.
  - **Avantages:** Hautement précises pour les cas conformes aux règles.
  - **Inconvénients:** Laborieuses à créer et à maintenir, ne s'adaptent pas bien aux variations linguistiques et aux nouvelles expressions.

Méthodes de POS Tagging

- **Méthodes Basées sur des Règles (Rule-based methods)**
  - **Prenons la phrase : "Le chat noir mange la souris."**
    - **Règle 1 :** Les mots qui se terminent par "Cons" sont généralement des noms (**N**).
      - **Chat** est un nom car il se termine par "t".
    - **Règle 2 :** Les mots qui décrivent des noms sont des adjectifs (**J**).
      - **Noir** est un adjectif car il décrit le nom "chat".
    - **Règle 3 :** Les mots qui expriment une action sont des verbes (**V**).
      - **Mange** est un verbe car il exprime l'action du chat.
    - **Règle 4 :** Les mots qui précèdent les noms sont des déterminants (**DT**).
      - **Le** est un déterminant car il précède le nom "chat".
  - En appliquant ces règles, nous obtenons la phrase suivante avec les POS tags :
    - **Le (DT) chat (N) noir (J) mange (V) la (DT) souris (N).**

Méthodes de POS Tagging

- **Méthodes Basées sur des Règles (Rule-based methods)**
  - **Avantages de cette méthode :**
    - **Facile à comprendre et à implémenter.**
    - **Hautement précise pour les cas conformes aux règles.**
  - **Inconvénients :**
    - **Laborieuse à créer et à maintenir les règles.**
    - **Ne s'adapte pas bien aux variations linguistiques et aux nouvelles expressions.**
  - **Exemple de Règle plus complexe :**
    - Si un mot est suivi d'un adjectif et d'un nom, il est probablement un verbe.
    - **Exemple :** "Le chat **court vite**".

Méthodes de POS Tagging

- **Méthodes Basées sur des Règles (Rule-based methods)**

```
def rule_based_pos_tagger(text):
  """
  This function takes a text string and assigns basic POS tags using rules.

  Args:
      text: A string containing the text to be tagged.

  Returns:
      A list of tuples, where each tuple contains a word and its POS tag.
  """
  tokens = text.split()
  tagged_words = []
  for token in tokens:
    # Rule 1: Words ending in "ing" are li
    if token.endswith("ing"):
      tagged_words.append((token, "VB"))
    # Rule 2: Words ending in "ly" are lik
    elif token.endswith("ly"):
      tagged_words.append((token, "RB"))
    # Rule 3: Words containing digits are likely numbers (CD)
    elif any(char.isdigit() for char in token):
      tagged_words.append((token, "CD"))
    # Rule 4: All other words are assumed to be nouns (NN)
    else:
      tagged_words.append((token, "NN"))
  return tagged_words

# Example usage
text = "This course is currently unavailable online"
tagged_words = rule_based_pos_tagger(text)

for word, tag in tagged_words:
  print(f"{word} ({tag})")
```

```
This (NN)
course (NN)
is (NN)
currently (RB)
unavailable (NN)
online (NN)
```

Méthodes de POS Tagging

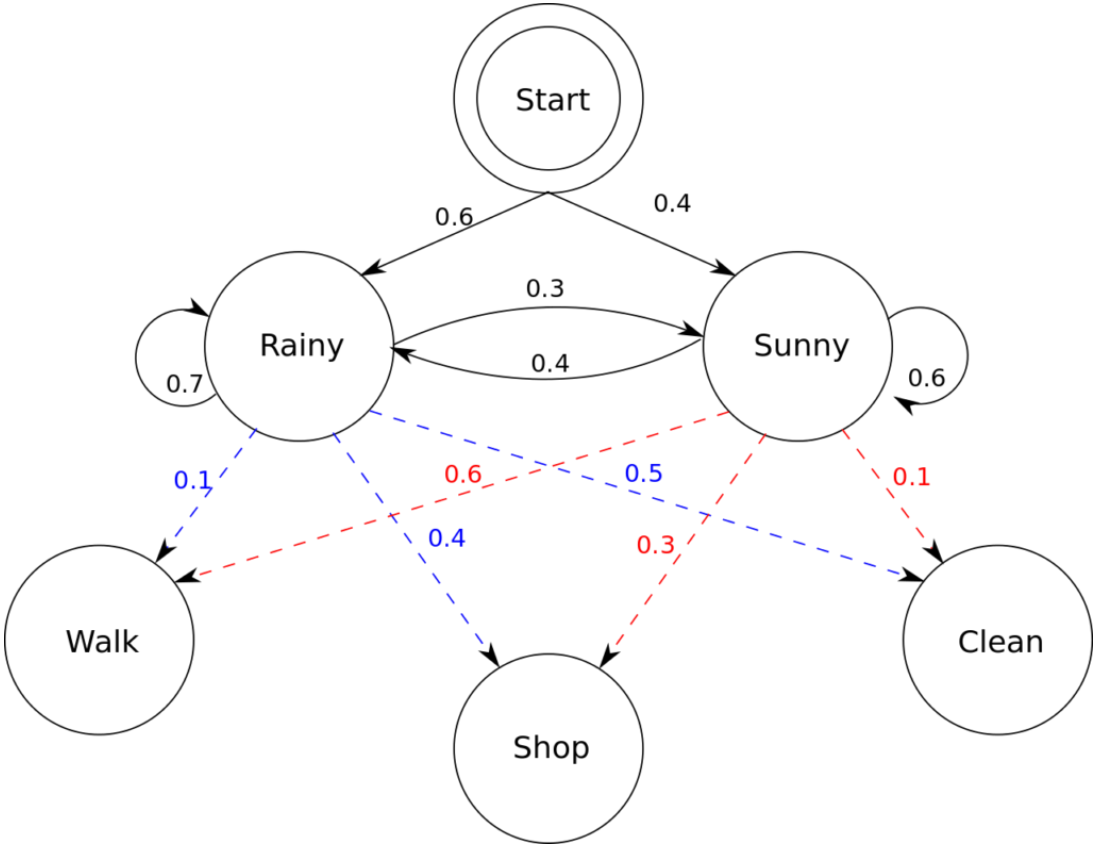
- **Méthodes Statistiques (Statistical methods)**
  - Ces méthodes basées sur des règles, les méthodes statistiques s'appuient sur la **probabilité** et l'**apprentissage automatique** pour étiqueter les mots.
  - **Collecte de données:**
    - Un grand corpus de textes déjà étiquettés (POS taggés) est constitué.
    - Ce corpus sert de base d'apprentissage pour l'algorithme.
  - **Entraînement:**
    - L'algorithme analyse le corpus et apprend les relations statistiques entre les mots, leur contexte et leur catégorie POS la plus probable.
    - Il prend en compte des facteurs comme la morphologie du mot, sa position dans la phrase, et le POS des mots voisins.

Méthodes de POS Tagging

- **Méthodes Statistiques (Statistical methods)**
  - **Étiquetage automatique:**
    - Une fois l'algorithme entraîné, il peut être utilisé pour étiqueter automatiquement les mots d'une nouvelle phrase.
    - L'algorithme calcule la probabilité de chaque catégorie POS pour chaque mot et choisit la plus probable.
  - **Avantages**
    - Plus flexibles et adaptables
    - Nécessitent moins d'intervention humaine
    - Peuvent être optimisées
  - **Inconvénients des méthodes statistiques :**
    - Nécessitent de grands corpus d'entraînement
    - Peuvent être moins précises
    - Difficiles à interpréter

Méthodes de POS Tagging

- **Méthodes Statistiques (Statistical methods)**



Méthodes de POS Tagging

- **Méthodes Statistiques (Statistical methods)**

```python
# Example text
text = "This course is currently unavailable online"
# Statistical tagging using NLTK (requires installation)
import nltk
from nltk import word_tokenize, pos_tag

# Download necessary resources (comment out if already downloaded)
nltk.download('punkt')

# Statistical tagging
nltk_tokens = word_tokenize(text)
nltk_tags = pos_tag(nltk_tokens)

print("Statistical Tags (NLTK):")
for word, tag in nltk_tags:
  print(f"{word} ({tag})")
```

```
Statistical Tags (NLTK):
This (DT)
course (NN)
is (VBZ)
currently (RB)
unavailable (JJ)
online (NN)
```

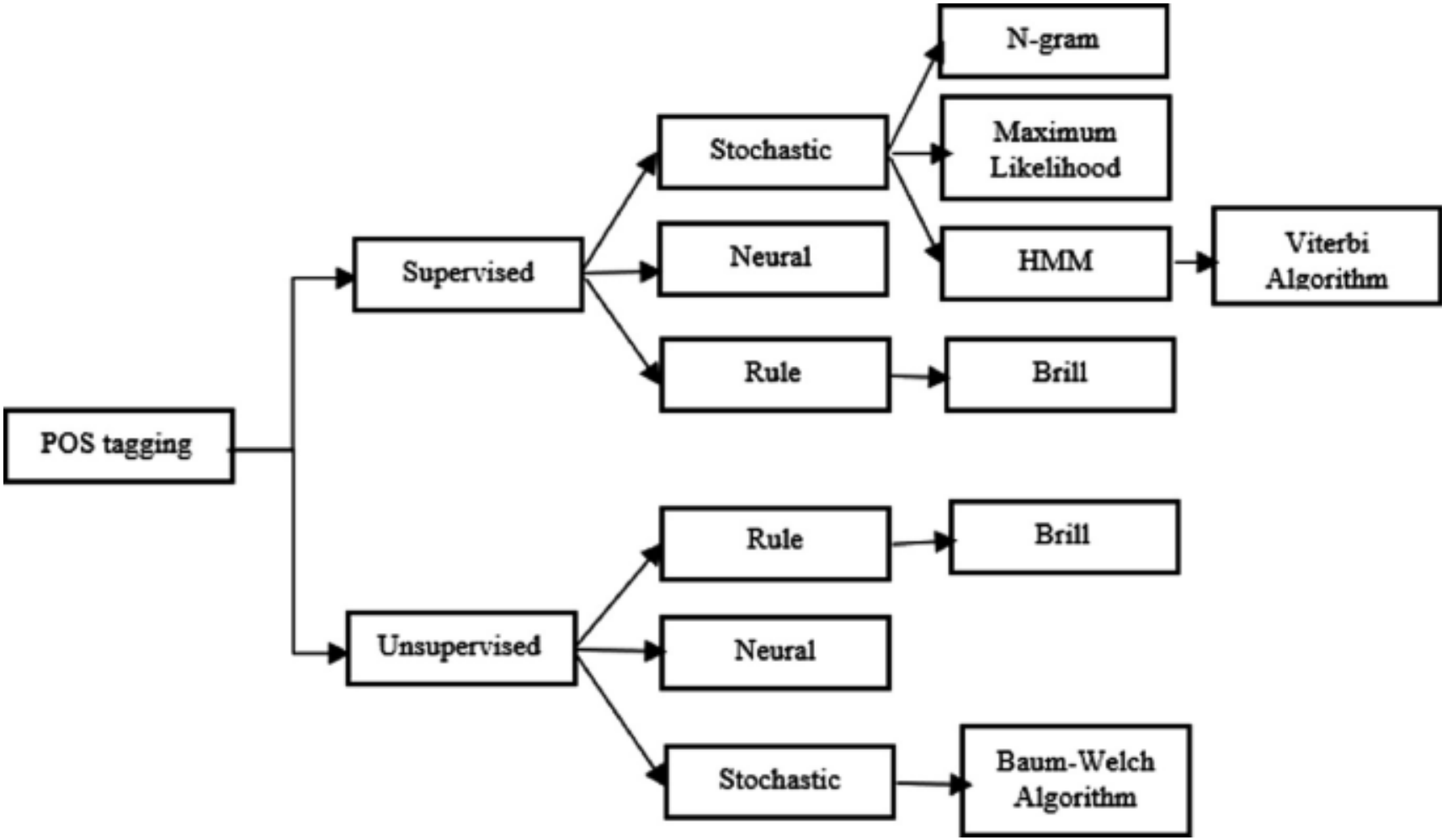
Méthodes de POS Tagging

- **Méthodes Hybrides pour le POS Tagging**
  - Les méthodes hybrides combinent les forces des approches basées sur des règles et des méthodes statistiques pour le POS Tagging.
  - **Avantages des méthodes hybrides :**
    - **Flexibilité et adaptabilité accrues:** Elles peuvent s'adapter à différents types de textes et de langues.
    - **Meilleure précision:** Elles exploitent les avantages des deux approches pour obtenir des résultats plus précis.
    - **Robustesse accrue:** Elles sont moins sensibles aux erreurs et aux variations du langage.

Méthodes de POS Tagging

- **Méthodes Hybrides pour le POS Tagging**
  - **Fonctionnement des méthodes hybrides :**
    - **Deux modules principaux:**
      - **Module basé sur des règles:** Applique des règles linguistiques pour identifier les POS des mots.
      - **Module statistique:** Utilise des modèles statistiques pour affiner les prédictions du module basé sur des règles.
    - **Différents types d'intégration :**
      - **Séries:** Le module statistique corrige les prédictions du module basé sur des règles.
      - **Parallèle:** Les deux modules font des prédictions indépendantes qui sont ensuite combinées.
      - **Imbriquée:** Le module statistique est utilisé pour sélectionner les règles les plus appropriées à appliquer.

Méthodes de POS Tagging



Évaluation du POS Tagging

- L'évaluation du POS Tagging est cruciale pour mesurer la performance d'un système et identifier d'éventuelles zones d'amélioration. Voici quelques métriques couramment utilisées :
  - **Précision (Accuracy):**
    - Calcule le pourcentage de mots correctement étiquetés par rapport au nombre total de mots.
    - **Formule :** Précision = (Mots correctement étiquetés) / (Nombre total de mots)
  - **Rappel (Recall):**
    - Indique la capacité du système à identifier tous les mots d'une catégorie POS spécifique.
    - Calcule le pourcentage de mots d'une catégorie donnée qui ont été correctement identifiés.
    - **Formule :** Rappel = (Mots d'une catégorie identifiés correctement) / (Nombre total de mots dans cette catégorie)

Évaluation du POS Tagging

- **F1-Score:**
  - Combinaison harmonique de la précision et du rappel.
  - Offre un équilibre entre les deux métriques.
  - **Formule :** F1-Score = 2 * (Précision * Rappel) / (Précision + Rappel)
- **Taux d'erreur (Error Rate):**
  - Pourcentage de mots mal étiquetés par le système.
  - **Formule :** Taux d'erreur = 1 - Précision
- **Matrices de Confusion:**
  - Représentation visuelle de la performance du système.
  - Montre comment les mots ont été prédits par rapport à leur catégorie POS réelle.
  - Utile pour identifier les types d'erreurs les plus fréquents.

Défis et Limitations du POS Tagging

- Le POS Tagging est une technique incontournable en NLP, mais il n'est pas exempt de défis et de limitations. Voici quelques points à garder à l'esprit :
  - **Ambiguïté :**
    - Certains mots peuvent appartenir à plusieurs catégories POS en fonction du contexte (ex : "**Pierre court**" vs "**Pierre Curie**").
    - Les systèmes de POS Tagging peuvent commettre des erreurs en cas d'ambiguïté.
  - **Taille et Qualité du Corpus d'entraînement :**
    - Les méthodes statistiques dépendent fortement de la taille et de la qualité du corpus d'entraînement utilisé.
    - Un corpus limité ou biaisé peut conduire à des performances médiocres, en particulier pour des domaines spécifiques.

Défis et Limitations du POS Tagging

- **Complexité des Langues :**
  - Certaines langues présentent des structures grammaticales plus complexes que d'autres, ce qui peut rendre le POS Tagging plus difficile.
  - Les langues agglutinantes (ex : finnois) où les mots sont formés par l'ajout de suffixes peuvent être plus complexes à analyser pour le POS Tagging.
- **Ressources et Coût :**
  - Développer et maintenir des systèmes de POS Tagging performants nécessite des ressources importantes, en particulier pour les langues moins courantes.

Défis et Limitations du POS Tagging

- **Amélioration des algorithmes:** Les chercheurs développent constamment de nouvelles techniques pour mieux gérer l'ambiguïté et les variations linguistiques.
- **Utilisation de corpus enrichis:** L'intégration de dictionnaires et de ressources linguistiques supplémentaires peut améliorer la performance.
- **Adaptation par domaine:** Entraîner des modèles spécifiques à des domaines particuliers (ex : juridique, médical) peut accroître la précision.
- **Supervision humaine:** La révision et la correction par des humains restent importantes pour garantir la qualité des résultats.

Exemple

```
]:  #!pip install spacy
    #!python -m spacy download en_core_web_lg
    #!python -m spacy download en_core_web_sm
```

```
]:  import spacy
    import en_core_web_sm

    #nlp = en_core_web_sm.load()
    nlp = spacy.load("en_core_web_sm")
    doc1 = nlp("Hassan Tower is located in Rabat")
    for word in doc1:
        print(word.text, word.pos_)
```

```
Hassan PROPN
Tower PROPN
is AUX
located VERB
in ADP
Rabat PROPN
```

Example of POS Tagging

Consider the sentence: “The quick brown fox jumps over the lazy dog.”

After performing POS Tagging:

- “The” is tagged as determiner (DT)
- “quick” is tagged as adjective (JJ)
- “brown” is tagged as adjective (JJ)
- “fox” is tagged as noun (NN)
- “jumps” is tagged as verb (VBZ)
- “over” is tagged as preposition (IN)
- “the” is tagged as determiner (DT)
- “lazy” is tagged as adjective (JJ)
- “dog” is tagged as noun (NN)