Université Mohammed V- Rabat
Ecole Mohammadia d'Ingénieurs
Département Génie Informatique

Filière Génie Informatique et Digitalisation

Text Analytics

Dr. H. SEBBAQ *h.sebbaq@gmail.com*

Pr. N. EL FADDOULI *nfaddouli@gmail.com*

2023-2024

NLP : Niveaux - Analyse sémantique

- L'analyse sémantique peut être divisée en deux parties, comme suit :
  - L'étude du sens du **mot individuel** est appelée **sémantique lexicale**.
  - L'étude de la façon dont les mots individuels se combinent pour fournir **un sens dans les phrases ou les paragraphes**, dans **le contexte du traitement** d'une unité plus large de langage naturel.

- **Exemple :** 'la maison blanche est superbe'
  - Signifier que l'affirmation est en contexte avec la Maison Blanche aux États-Unis
  - Affirmation parle littéralement d'une maison à proximité, dont la couleur est blanche et qui est super.

→Comprendre le sens approprié de la phrase est la tâche de l'analyse sémantique.

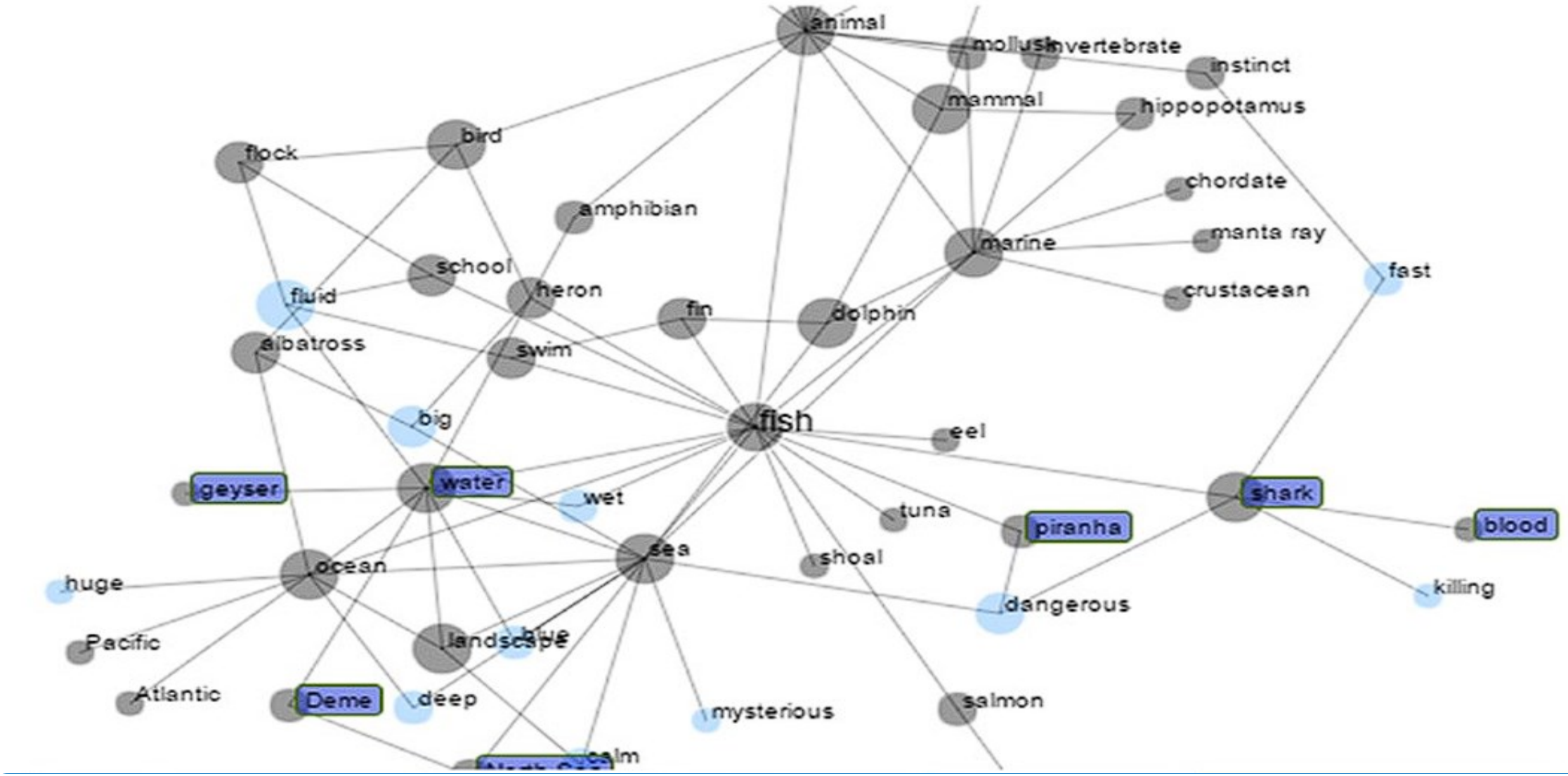
L'analyse sémantique s'intéresse à la représentation du sens. Elle se concentre principalement sur le sens littéral des mots, des expressions et des phrases.

L'analyseur sémantique ne tient pas compte d'une phrase telle que **"hot ice-cream".**

NLP : Niveaux - Analyse sémantique – Sémantique lexicale

- **Etudie les relations sémantiques entre les unités lexicales dans une langue.**
- Les unités lexicales, représentées par des morphèmes, sont les plus petits éléments porteurs de sens dans une langue.
- Chaque unité lexicale a sa propre syntaxe, forme et signification, influencée par son contexte dans des phrases, des clauses et des phrases.
- Un lexique regroupe l'ensemble de ces unités lexicales.
- Les **lemmes** sont les formes canoniques ou citations d'un ensemble de mots, tandis que les formes de mots sont des formes fléchies du lemme.
- **Les homonymes, homographes et homophones** désignent respectivement des mots avec des significations, des orthographes ou des prononciations similaires ou différentes.
- **Les hétéronymes et hétérographes** sont des mots ayant une orthographe identique mais une prononciation ou une signification différente.
- **Les polysèmes** ont des significations différentes mais liées, tandis que les **capitonyms** changent de sens selon s'ils sont écrits en majuscules ou en minuscules.
- **Les synonymes** partagent des significations similaires, alors que les antonymes sont des mots opposés.
- **Les hyponymes** sont des mots spécifiques inclus dans des catégories plus larges appelées hyperonymes, représentant des mots plus génériques.

NLP : Niveaux - Analyse sémantique – Réseaux sémantiques



NLP : Niveaux - Analyse sémantique – Réseaux sémantiques

- Les réseaux sémantiques représentent les concepts et leurs relations via des graphes.
- Chaque concept est un nœud du graphe, et les relations entre eux sont des arêtes orientées ou non orientées.
- Les réseaux sémantiques utilisent des relations telles que « is-a", « has-a", « is-part-of", 'is-related-to' etc.
- Ces réseaux modélisent les connaissances en reliant des concepts et permettent d'explorer les relations entre eux.
- Le web sémantique étend le web en utilisant des métadonnées et des techniques de modélisation de données pour créer une représentation plus sémantique du contenu en ligne.
- **WordNet est une riche base de données lexicale utilisant des synsets (ensembles de synonymes) pour représenter les relations sémantiques entre les mots.**

IS-A	HAS-SPECIFIC	IS-PART-OF	HAS-PART	HAS-PROPERTY	IS-RELATED-TO
animal	eel	ocean	fin	wet	heron
marine	piranha	school			
	salmon	sea			
	shark	water			
	shoal				
	tuna				

NLP : Niveaux - Analyse sémantique - WordNet

http://wordnetweb.princeton.edu/perl/webwn

	Synset	Part of Speech	Definition	Lemmas	Examples
0	Synset('school.n.01')	noun.group	an educational institution	[school]	[the school was founded in 1900]
1	Synset('school.n.02')	noun.artifact	a building where young people receive education	[school, schoolhouse]	[the school was built in 1932, he walked to school every morning]
2	Synset('school.n.03')	noun.cognition	the process of being formally educated at a school	[school, schooling]	[what will you do when you finish school?]
3	Synset('school.n.04')	noun.group	a body of creative artists or writers or thinkers linked by a similar style or by similar teachers	[school]	[the Venetian school of painting]
4	Synset('school.n.05')	noun.time	the period of instruction in a school; the time period when school is in session	[school, schooltime, school_day]	[stay after school, he didn't miss a single day of school, when the school day was done we would walk home together]
5	Synset('school.n.06')	noun.group	an educational institution's faculty and students	[school]	[the school keeps parents informed, the whole school turned out for the game]
6	Synset('school.n.07')	noun.group	a large group of fish	[school, shoal]	[a school of small glittering fish swam by]
7	Synset('school.v.01')	verb.social	educate in or as if in a school	[school]	[The children are schooled at great cost to their parents in private institutions]
8	Synset('educate.v.03')	verb.social	teach or refine to be discriminative in taste or judgment	[educate, school, train, cultivate, civilize, civilise]	[Cultivate your musical taste, Train your tastebuds, She is well schooled in poetry]
9	Synset('school.v.03')	verb.motion	swim in or form a large group of fish	[school]	[A cluster of schooling fish was attracted to the bait]

NLP : Niveaux - Analyse sémantique - Représentation de la sémantique

- La représentation de la sémantique est particulièrement utile pour effectuer diverses opérations de traitement du langage naturel afin de permettre aux machines de comprendre la sémantique des messages à l'aide de représentations appropriées, étant donné qu'elles n'ont pas le pouvoir cognitif de l'homme.

NLP : Niveaux - Analyse sémantique - Représentation de la sémantique

- **La logique propositionnelle concerne les propositions et leurs connexions logiques, représentées par des symboles et des connecteurs. Ce cadre logique permet d'interpréter la signification des propositions en utilisant des règles logiques et d'exprimer les relations sémantiques entre elles.**

p : "Il pleut." ***q*** : "Je prends mon parapluie."

Nous pouvons utiliser des opérateurs logiques pour former des combinaisons de ces propositions :

1.Conjonction (∧, ET) → "Il pleut et je prends mon parapluie." → $p \wedge q$

2.Disjonction (V, OU) → "Il pleut ou je prends mon parapluie. " → $p \vee q$

3.Négation (¬, NON) : "Il ne pleut pas." → $\neg p$

4.Implication (→, SI...ALORS) : "S'il pleut, alors je prends mon parapluie." →:$p \rightarrow q$

5.Équivalence (↔, SI ET SEULEMENT SI) : "Je prends mon parapluie si et seulement s'il pleut." → $p \leftrightarrow q$

NLP : Niveaux - Analyse sémantique - Représentation de la sémantique

- La logique du premier ordre étend la logique propositionnelle en introduisant des variables, des quantificateurs (comme ∀ et ∃), des prédicats pour décrire les relations entre objets spécifiques, des fonctions, et des constantes. Elle permet des affirmations détaillées sur des ensembles d'objets en utilisant des formules logiques soumises à des règles syntaxiques et sémantiques précises

Considérons un domaine de discours où nous avons des individus, disons des personnes, et une relation binaire "estParentDe(x, y)" qui représente la relation parentale entre deux personnes, où x est le parent de y.

Maintenant, exprimons quelques assertions avec la logique du premier ordre :

- Assertion : "Pour tout x, s'il existe un y tel que x est parent de y, alors x est un parent.

  " Formulation logique : $\forall x \exists y \, estParentDe(x,y) \rightarrow Parent(x)$

- Assertion : "Il existe une personne x qui est un parent de quelqu'un.

  " Formulation logique : $\exists x \exists y estParentDe(x,y)$

NLP : Niveaux - Analyse sémantique - Représentation de la sémantique

- **Représentations vectorielles (embedding) :** Ces représentations transforment les mots ou les phrases en vecteurs numériques dans un espace mathématique. Des méthodes telles que Word Embeddings (comme Word2Vec, GloVe, FastText), Transformer-based (BERT, GPT, etc.) sont utilisées pour représenter le sens des mots ou des phrases sous forme de vecteurs numériques.

- **Réseaux de neurones récurrents (RNN) et réseaux de neurones récurrents à mémoire à court et long terme (LSTM) :** Ces modèles utilisent des architectures neuronales pour capturer les dépendances séquentielles et sémantiques dans les phrases ou les documents.

- **Réseaux de neurones convolutionnels (CNN) :** Ces modèles sont souvent utilisés pour extraire des caractéristiques sémantiques des données textuelles, en particulier pour des tâches comme la classification de texte ou l'extraction d'informations.

- **Transformers :** Ces modèles, comme BERT (Bidirectional Encoder Representations from Transformers) ou GPT (Generative Pretrained Transformer), sont des architectures de réseaux neuronaux révolutionnaires qui ont permis d'obtenir des représentations sémantiques très performantes en exploitant des mécanismes d'attention

NLP : Niveaux - Analyse pragmatique

- Traite des connaissances extérieures aux mots, ce qui signifie des connaissances externes aux documents et/ou requêtes.
- Se concentre sur ce qui a été décrit est réinterprétée par ce qu'elle signifiait réellement, en tirant les différents aspects du langage qui nécessitent une connaissance du monde réel.
- Exemple :
  - ‘Pruning a tree is a long process’.
  - "Pruning" est un concept dans les techniques algorithmiques en informatique, pas lié à la taille d'un arbre réel, mais à des processus informatiques.
  - Cette distinction crée une ambiguïté, et la façon de gérer de telles situations est un domaine de recherche ouvert.
  - Les grandes entreprises technologiques utilisent des méthodes d'apprentissage approfondi pour réaliser une analyse pragmatique et comprendre le contexte exact des phrases.
  - L'objectif est de développer des applications de traitement du langage naturel (NLP) très précises en utilisant cette analyse contextuelle.

NLP : Niveaux - Analyse du discours

L'intégration du discours est étroitement liée à la pragmatique.
Cette analyse porte sur la manière dont la phrase immédiatement précédente peut affecter le sens et l'interprétation de la phrase suivante. Ici, le contexte peut être analysé dans un contexte plus large, comme le niveau du paragraphe, le niveau du document, etc.

Elle traite de la manière dont la phrase immédiatement précédente peut affecter l'interprétation de la phrase suivante.

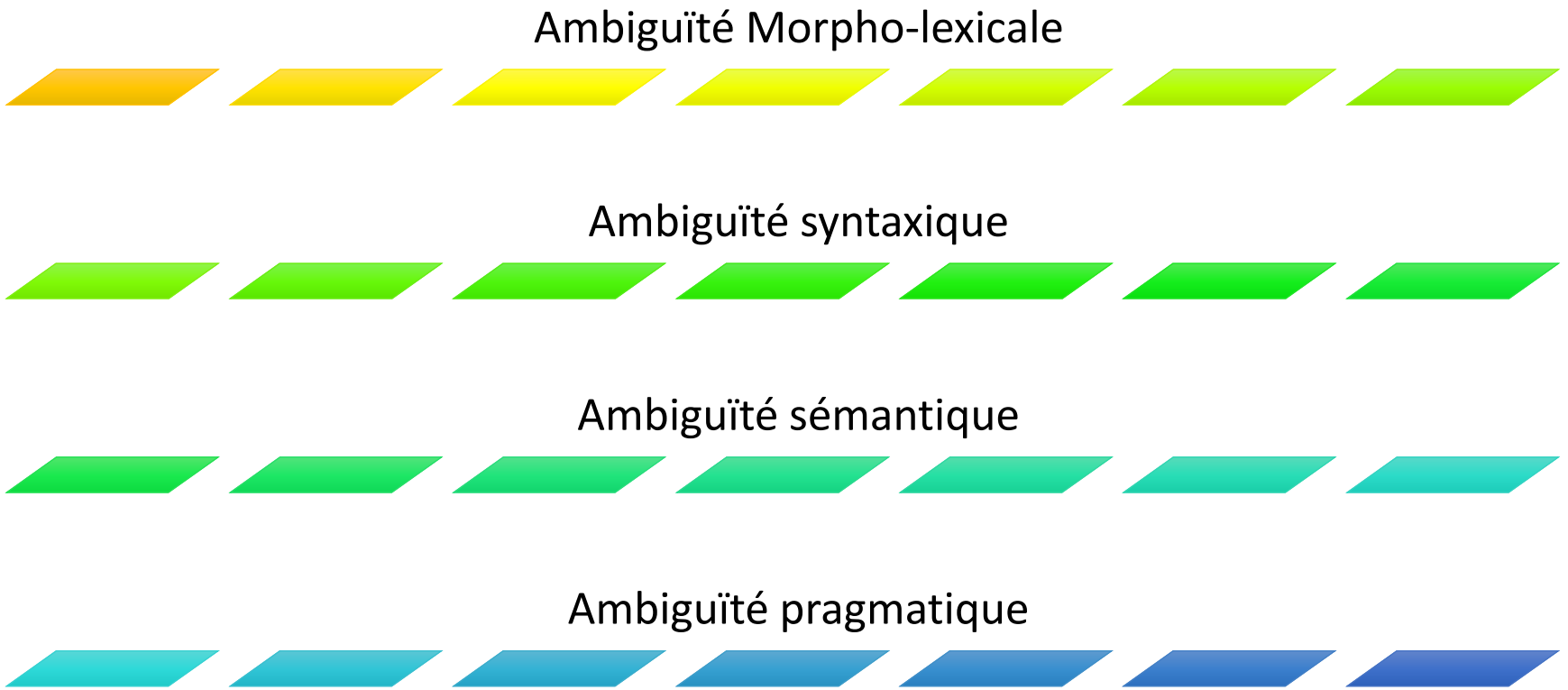
Rabat (en arabe : الرباط, *ar-Ribāṭ* ; en amazighe : ⵔⵔⴱⴰⵟ[5], *Rrbaṭ* ; en arabe marocain : الرّبّاط, *er-Rbaṭ*) est la capitale du Maroc. La ville est située au bord de l'Atlantique au nord-ouest du Maroc, à 40 km au sud de Kénitra et 240 km au sud-ouest de Tanger et du détroit de Gibraltar, et à 87 km au nord-est de Casablanca. Elle est séparée de la ville de Salé au niveau de l'embouchure du Bouregreg, d'où leur surnom de « villes jumelles »[6].

NLP : Problème de l’ambiguïté

- Lorsque nous nous lançons dans l'analyse sémantique, nous pouvons constater que de nombreux cas sont trop ambigus pour être traités par un système de NLP.
- Dans ces cas, nous devons savoir quels types d'ambiguïté existent et comment les traiter.
- L'ambiguïté est l'un des domaines du NLP et des sciences cognitives qui n'a pas de solution bien définie.
- Parfois, les phrases sont si complexes et ambiguës que seul le locuteur peut en définir le sens original ou définitif.
- Un mot, une expression ou une phrase est ambiguë si elle a plus d'un sens.
- Le mot ‘Light’, il peut signifier "pas très lourd" ou "pas très sombre". Il s'agit d'une ambiguïté au niveau du mot.
- L'expression "récipient à œufs en porcelaine" est une ambiguïté au niveau de la structure.

NLP : Problème de l'ambiguïté

les différents types d'ambiguïté



NLP : Problème de l'ambiguïté

Comment déterminer le genre du mot *livre* dans les phrases suivantes :

- **J'ai lu un livre**
- **Il ne s'agit pas de livres mais de lires**

→ par un traitement morphologique ?

Pour la première phrase, il faut repérer que *livre* est précèdé de l'article *un*

→ traitement syntaxique !

Pour la seconde, il faut intégrer des connaissances sur le monde et la situation de communication (*livre* et *lire* sont deux monnaies)

→ traitement pragmatique ! !

Un mot/une phrase peut avoir plusieurs significations.

– Fall ● The third season of the year ● Moving down towards the ground or towards a lower position

– The door is open. ● Expressing a fact ● A request to close the door

NLP : Problème Ambiguïté – Analyse Morpho-lexicale

L'ambiguïté lexicale est une ambiguïté au niveau des mots. Un seul mot peut avoir une signification ambiguë en de sa structure interne et de sa classe syntaxique. Prenons quelques exemples :

Look at the stars → Verbe.

The person gave him a warm **look**. → Nom

She won three **silve**r medals. → Nom

She made **silver** speech.→ Adjectif

His stress had **silvered** his hair. → Verbe

Il a été **averti** du danger → verbe

Un consommateur **averti** → adjectif

Dans ces exemples, des mots spécifiques changent d'étiquette POS en fonction de leur utilisation dans la structure de la phrase.

Ce type d'ambiguïté peut être résolu en utilisant deux approches :

- En utilisant des outils d'étiquetage POS précis
- Un sens dans WordNet comporte différentes significations disponibles pour un mot lorsque ce dernier est associé à des catégories grammaticales spécifiques. Cela aide également à gérer l'ambiguïté.

NLP : Problème de l'ambiguïté – Analyse syntaxique

I saw the man with a telescope.
– Who had the telescope?

Rapport de vraisemblance : Utiliser des approches statistiques et obtenir le rapport de vraisemblance le plus élevé. Nous devons examiner les co-occurrences entre le verbe et la préposition d'une part, et la préposition et le nom d'autre part, puis calculer le rapport de vraisemblance logarithmique en utilisant l'équation suivante :

$$F(v,n,p) = log\frac{p(p/v)}{p(p/n)}$$

p(p/v) représente la probabilité de voir un groupe prépositionnel (PP) avec la préposition p après le verbe v.
p(p/n) est la probabilité de voir un PP avec la préposition p après le nom n.

Si F(v,p,n) < 0, alors nous devons attacher la préposition au nom, et si F(v,p,n) > 0, alors nous devons attacher la préposition au verbe.

NLP : Problème de l'ambiguïté – Analyse sémantique

The astronomer loves the star.
– Star in the sky
– Celebrity

- Gestion de l'ambiguïté sémantique : Un domaine de recherche ouvert dans le traitement du langage naturel.
- Word Sense Disambiguation (WSD) : Algorithme de Lesk
- Word2Vec

NLP : Problème de l'ambiguïté – Analyse sémantique

```python
import nltk
nltk.download('punkt')
from nltk.wsd import lesk
from nltk import word_tokenize

# sample text and word to disambiguate
samples = [('The fruits on that plant have ripened', 'n'),
           ('He finally reaped the fruit of his hard work as he won the race', 'n')]

# perform word sense disambiguation
word = 'fruit'
for sentence, pos_tag in samples:
    word_syn = lesk(word_tokenize(sentence.lower()), word, pos_tag)
    print('Sentence:', sentence)
    print('Word synset:', word_syn)
    print('Corresponding defition:', word_syn.definition())
    print()
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
Sentence: The fruits on that plant have ripened
Word synset: Synset('fruit.n.01')
Corresponding defition: the ripened reproductive body of a seed plant

Sentence: He finally reaped the fruit of his hard work as he won the race
Word synset: Synset('fruit.n.03')
Corresponding defition: the consequence of some effort or action
```

NLP : Problème de l'ambiguïté – Analyse du discours

- Alice understands that you like your mother, but she …

– 'She' se réfère a la maman ou bien Alice?

NLP : Problème de l’ambiguïté – Analyse pragmatique

"Il a besoin de lunettes."

Interprétation 1 : Cette personne a besoin de lunettes pour voir correctement.
Interprétation 2 : Cette personne a besoin de lunettes en tant qu'accessoire de mode.

La signification de la phrase peut varier en fonction du contexte et des indices donnés par la situation ou la conversation, créant ainsi une ambiguïté pragmatique.