

Programmation de Modèles Linguistiques (I) (TALA330A L3 INALCO)

Crédits : Karën Fort (pépites pour le TAL)

Caroline Parfait

caroline.parfait@sorbonne-universite.fr

2021-2022

Obtic, Sorbonne Center for Artificial Intelligence
STIH EA 4509, Sorbonne Université

Plan de la présentation

1. Introduction
2. Modélisation de la Langue et TAL
3. En pratique pour la machine ?

Organisation du semestre

- CM + TD le Lundi de 16h30 à 19h30 (C. Parfait)
- Évaluation : Questions de cours + Exercice de programmation

Ressources

- TAL et Linguistique Informatique 1, ISTE Ed. (Mohamed Z. Kurdi)
- Speech and Language Processing (Dan Jurafsky),
<https://web.stanford.edu/~jurafsky/slp3/>
- Helpdesk : mail ou bureau 206/212 à Serpente (sur RDV)

Introduction

Modélisation de la Langue et TAL

- Le TAL en 2 minutes

- Le TAL vu par ses applications

- Le TAL : définitions

- Ambiguïté à tous les étages

- Étapes de traitement

En pratique pour la machine ?

Introduction

Contexte : l'accès aux données langagières



Difficulté : des données conçues pour des humains



Solution : des données sous-traitées à des machines



Modéliser la langue

- Observables : le caractère, le phonème, le token
- Eco-système : n-grammes, sons, proposition/phrased
- Applications : identification de langue, d'auteur, de style ; indexation et génération de texte ; extraction d'information

Modélisation de la Langue et TAL

- Traitement Automatique des Langues
- Traitement Automatique du Langage Naturel

- Traitement Automatique des Langues
- Traitement Automatique du Langage Naturel
- *Natural Language Processing*
- Technologies du Langage (Humain)
- Sciences de l'information
- Linguistique Computationnelle

Reoiajr oj earoij reoa o eo ao aeoi oj aroij aoeir eoaj.

Langue inconnue

La langue vue par la machine

Reoiajr oj earoij reoa o eo ao aeoi oj aroij aeir eoaj.

Découper

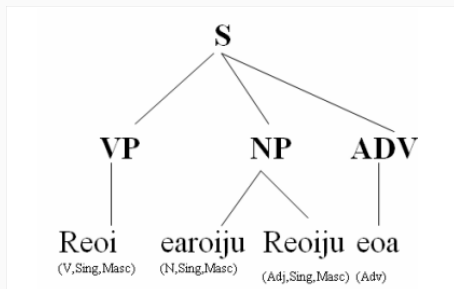
La langue vue par la machine

Reoi earoiju Reoiju eoa
(V,Sing,Masc) (N,Sing,Masc) (Adj,Sing,Masc) (Adv)

Reoiajr oj earoij reoa o eo ao aeoi oj aeroij aoeir eoaj.

Etiqueter

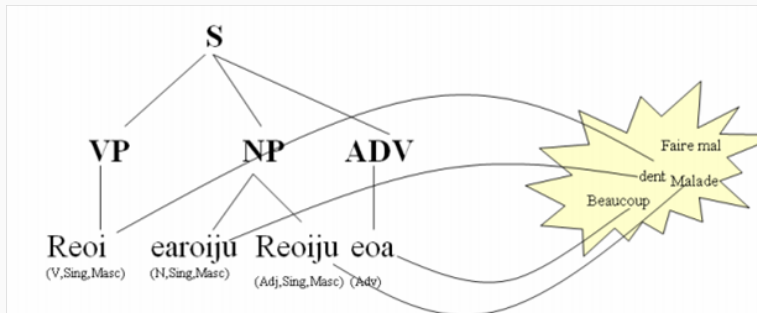
La langue vue par la machine



Reoiajr oj earoij reoa o eo ao aeoi oj aeroij aeir eoaj.

Reconnaître la structure

La langue vue par la machine



Reoiajr oj earoij reoa o eo ao aeoi oj aeroij aoeir eoaj.

Évaluer le sense et mettre en contexte

Nécessité d'explicitation (réductionnisme)

Pour quoi faire ?

- Traduction automatique
- Recherche d'information
- Indexation
- Traitement d'Emails
- Veille
- E-reputation
- Big Data
- ...

Pour les chercheurs :

- comprendre comment fonctionne la langue
- exemplifier/vérifier des théories

Pour les industriels :

- reproduire l'activité langagière
- optimiser les performances des systèmes
- offrir de nouveaux services

Introduction

Modélisation de la Langue et TAL

- Le TAL en 2 minutes

- Le TAL vu par ses applications

- Le TAL : définitions

- Ambiguïté à tous les étages

- Étapes de traitement

En pratique pour la machine ?

- *La mousse aux fraises est sur la table de l'avocat.*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant. . .*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant. . .*
 - *. . .je l'ai entendu freiner.*

Le TAL : pas si simple. . .

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant. . .*
 - *. . .je l'ai entendu freiner.*
 - *. . .je l'ai entendu crier.*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant...*
 - *...je l'ai entendu freiner.*
 - *...je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur...*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant...*
 - *...je l'ai entendu freiner.*
 - *...je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur...*
 - *...il faisait trop de bruit.*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant...*
 - ...je l'ai entendu freiner.
 - ...je l'ai entendu crier.
- *Le professeur a envoyé l'élève chez le proviseur...*
 - ...il faisait trop de bruit.
 - ...il était excédé.

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant...*
 - *...je l'ai entendu freiner.*
 - *...je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur...*
 - *...il faisait trop de bruit.*
 - *...il était excédé.*
 - *...il l'avait convoqué.*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer!*
- *Le bus a renversé un passant...*
 - ...je l'ai entendu freiner.
 - ...je l'ai entendu crier.
- *Le professeur a envoyé l'élève chez le proviseur...*
 - ...il faisait trop de bruit.
 - ...il était excédé.
 - ...il l'avait convoqué.

→ **ambiguïtés** pour les systèmes et/ou pour les humains

→ **inférences** difficiles à recenser exhaustivement

Le découpage en "mots" ou tokenization

- *l'arbre*

Le découpage en "mots" ou tokenization

- *l'arbre*
- *aujourd'hui*

Le découpage en "mots" ou tokenization

- *l'arbre*
- *aujourd'hui*
- *plateforme*

Le découpage en "mots" ou tokenization

- *l'arbre*
- *aujourd'hui*
- *plateforme*
- *plate-forme*

Le découpage en "mots" ou tokenization

- *l'arbre*
- *aujourd'hui*
- *plateforme*
- *plate-forme*
- *dit-il*

Le découpage en "mots" ou tokenization

- *l'arbre*
- *aujourd'hui*
- *plateforme*
- *plate-forme*
- *dit-il*
- *cé-*
sure

Le découpage en "mots" ou tokenization

- *l'arbre*
- *aujourd'hui*
- *plateforme*
- *plate-forme*
- *dit-il*
- *cé-*
sure

Le découpage en "mots" ou tokenization

- *l'arbre*
- *aujourd'hui*
- *plateforme*
- *plate-forme*
- *dit-il*
- *cé-*
sure
- Vablaheibárvégavinnuverkfærageymsluskúraútidyralklakkippuhringur
("Porte-clés de la chaîne de clé pour la porte extérieure du hangar à
outils des agents de la route sur le plateau Vablaheibi")

la *porte*

- *porte* +Nom +Fem + Sg

la *porte*

- *porte* + Nom + Fem + Sg
- *porte* + VT + 1/3P + Sg

Jean regarde un homme sur la colline avec un télescope.

- **Qui** est sur la colline ?
- **Qui** a un télescope ?
- Jean regarde [un homme sur la colline avec un télescope]

Jean regarde un homme sur la colline avec un télescope.

- **Qui** est sur la colline ?
- **Qui** a un télescope ?
- Jean regarde [un homme sur la colline avec un télescope]
- Jean regarde [un homme sur la colline] avec un télescope

Jean regarde un homme sur la colline avec un télescope.

- **Qui** est sur la colline ?
- **Qui** a un télescope ?
- Jean regarde [un homme sur la colline avec un télescope]
- Jean regarde [un homme sur la colline] avec un télescope
- Jean regarde un homme [sur la colline avec un télescope]

Jean regarde un homme sur la colline avec un télescope.

- **Qui** est sur la colline ?
- **Qui** a un télescope ?
- Jean regarde [un homme sur la colline avec un télescope]
- Jean regarde [un homme sur la colline] avec un télescope
- Jean regarde un homme [sur la colline avec un télescope]

Jean regarde un homme sur la colline avec un télescope.

- **Qui** est sur la colline ?
- **Qui** a un télescope ?
- Jean regarde [un homme sur la colline avec un télescope]
- Jean regarde [un homme sur la colline] avec un télescope
- Jean regarde un homme [sur la colline avec un télescope]

Un défi à l'intuition

Moi "... Vous avez oublié la pièce jointe..."

Le client mail :

2 mots-clés de pièce jointe trouvés : pièce jointe, jointe

[Ajouter une pièce jointe...](#)

[Me le rappeler plus tard](#)

Tous les hommes aiment une femme.

→ **Chaque** homme aime une femme ou

→ **tous** les hommes aiment la même femme ?

Tous les hommes aiment une femme.

→ **Chaque** homme aime une femme ou

→ **tous** les hommes aiment la même femme ?

The Wayward Pines Trilogy (2012–2014) is a mystery/thriller/science fiction novel series by American author Blake Crouch.

→ **Book** = "The Wayward Pines Trilogy", **Date** = "2012-2014",
Author = "Blake Crouch"

Tous les hommes aiment une femme.

→ **Chaque** homme aime une femme ou

→ **tous** les hommes aiment la même femme ?

The Wayward Pines Trilogy (2012–2014) is a mystery/thriller/science fiction novel series by American author Blake Crouch.

→ **Book** = "The Wayward Pines Trilogy", **Date** = "2012-2014",
Author = "Blake Crouch"

Jean-Marie Roughol, 47 ans, sort un livre écrit à quatre mains avec le président du Conseil constitutionnel, Jean-Louis Debré

→ **Book** = "un livre écrit à quatre mains", **Date** = "Today",
Author = "Jean-Marie Roughol, Président du Conseil Constitutionnel"

Différences de représentation et de combinaison

Русский язык	Français
первый	Premier
второй	Deuxième
этаж	Etage
??	Premier étage

Différences de représentation et de combinaison

Русский язык	Français
первый	Premier
второй	Deuxième
этаж	Etage
??	Premier étage
второй этаж	Premier étage
первый этаж	Rez-de-chaussée
резь-де-Chaussee	Rez-de-chaussée

En pratique pour la machine ?

Les difficultés de programmation algorithmique sur des données textuelles :

- Comment la machine reconnaît une chaîne de caractère ? un caractère ?
- Comment la machine reconnaît des phrases ? des mots ?