Résumé du CM2

Rappel : observables, tokens, lemmes, tri-grammes de caractères ...

Résumé du CM2

Rappel : observables, tokens, lemmes, tri-grammes de caractères ...

- **Systèmes** à base de règles VS systèmes à base de données
- Systèmes hybrides

Résumé du CM2

Rappel : observables, tokens, lemmes, tri-grammes de caractères ...

- **Systèmes** à base de règles VS systèmes à base de données
- Systèmes hybrides
- Loi de **Zipf** :
  - Les mots fréquents couvrent une grande partie des textes
  - Les mots rares sont très nombreux

Résumé du CM2

Rappel : observables, tokens, lemmes, tri-grammes de caractères ...

- **Systèmes** à base de règles VS systèmes à base de données
- Systèmes hybrides
- Loi de **Zipf** :
  - Les mots fréquents couvrent une grande partie des textes
  - Les mots rares sont très nombreux
- **Identification de Langue**
  - $\rightarrow$ Les mots vides ne portent pas de sens par eux-mêmes

Résumé du CM2

Rappel : observables, tokens, lemmes, tri-grammes de caractères ...

- **Systèmes** à base de règles VS systèmes à base de données
- Systèmes hybrides
- Loi de **Zipf** :
  - Les mots fréquents couvrent une grande partie des textes
  - Les mots rares sont très nombreux
- **Identification de Langue**
  - $\rightarrow$ Les mots vides ne portent pas de sens par eux-mêmes
- Proof of concept, **Limites** :

Résumé du CM2

Rappel : observables, tokens, lemmes, tri-grammes de caractères ...

- **Systèmes** à base de règles VS systèmes à base de données
- Systèmes hybrides
- Loi de **Zipf** :
  - Les mots fréquents couvrent une grande partie des textes
  - Les mots rares sont très nombreux
- **Identification de Langue**
  - $\rightarrow$ Les mots vides ne portent pas de sens par eux-mêmes
- Proof of concept, **Limites** :
  - Quand les textes sont courts
  - Quand les textes comportent plusieurs langues

Résumé du CM2

Rappel : observables, tokens, lemmes, tri-grammes de caractères ...

- **Systèmes** à base de règles VS systèmes à base de données
- Systèmes hybrides
- Loi de **Zipf** :
  - Les mots fréquents couvrent une grande partie des textes
  - Les mots rares sont très nombreux
- **Identification de Langue**
  - $\rightarrow$ Les mots vides ne portent pas de sens par eux-mêmes
- Proof of concept, **Limites** :
  - Quand les textes sont courts
  - Quand les textes comportent plusieurs langues
- **Évaluation** grossière/fine, quanti/quali

Résumé du CM2

Rappel : observables, tokens, lemmes, tri-grammes de caractères ...

- **Systèmes** à base de règles VS systèmes à base de données
- Systèmes hybrides
- Loi de **Zipf** :
  - Les mots fréquents couvrent une grande partie des textes
  - Les mots rares sont très nombreux
- **Identification de Langue**
  - $\rightarrow$ Les mots vides ne portent pas de sens par eux-mêmes
- Proof of concept, **Limites** :
  - Quand les textes sont courts
  - Quand les textes comportent plusieurs langues
- **Évaluation** grossière/fine, quanti/quali
- Précision VS bruit, Rappel Vs silence
  - F-mesure

Zoom sur la F-mesure

Évaluation combinée, la $F - mesure$: $F = 2 * \frac{P*R}{(P+R)}$

Moyenne **Harmonique** et non **Géométrique**

Zoom sur la F-mesure

Évaluation combinée, la $F - mesure$: $F = 2 * \frac{P*R}{(P+R)}$

Moyenne **Harmonique** et non **Géométrique**

	Précision	**R**appel	Moy. Géo	Moy. Harmo
Système 1	2	100	51	3,9
Système 2	98	4	51	7,7
Système 3	50	50	50	50,0
Système 4	60	40	50	48,0

Zoom sur la F-mesure

Évaluation combinée, la $F - mesure$: $F = 2 * \frac{P*R}{(P+R)}$

Moyenne **Harmonique** et non **Géométrique**

	Précision	**R**appel	Moy. Géo	Moy. Harmo
Système 1	2	100	51	3,9
Système 2	98	4	51	7,7
Système 3	50	50	50	50,0
Système 4	60	40	50	48,0

On peut aussi **pondérer** la F-mesure de manière à "favoriser" des systèmes bons en Rappel ou bons en Précision (β est un coefficient) :

Évaluation pondérée : $F_\beta = (1 + \beta^2) * \frac{P*R}{(\beta^2*P)+R}$

- $\beta = 1$: équilibrée, $\beta < 1$ favorise P (Précision) et inversement

Analyse syntaxique

Analyse syntaxique

Analyse syntaxique : le problème

Un énoncé est une suite ordonnée de mots

- Quels mots ?
- Dans quel(s) ordre(s) ?
- Pour quel sens ?

Analyse syntaxique : le problème

- Comment un locuteur sait-il qu'une phrase est bien formée ?

Analyse syntaxique : le problème

- Comment un locuteur sait-il qu'une phrase est bien formée ?
  - une tâche complexe
  - apprise par l'exemple...

Analyse syntaxique : le problème

- Comment un locuteur sait-il qu'une phrase est bien formée ?
  - une tâche complexe
  - apprise par l'exemple...
  - ... parfois sur des exemples faux ou fabriqués

Analyse syntaxique : le problème

- Comment un locuteur sait-il qu'une phrase est bien formée ?
  - une tâche complexe
  - apprise par l'exemple...
  - ... parfois sur des exemples faux ou fabriqués
- Des régularités dans l'acquisition du langage

Analyse syntaxique : le problème

- Comment un locuteur sait-il qu'une phrase est bien formée ?
  - une tâche complexe
  - apprise par l'exemple...
  - ... parfois sur des exemples faux ou fabriqués
- Des régularités dans l'acquisition du langage
  - Difficile après un certain âge (processus biologique, psychologique)
  - On peut apprendre n'importe quelle langue

Analyse syntaxique : le problème

- Comment un locuteur sait-il qu'une phrase est bien formée ?
  - une tâche complexe
  - apprise par l'exemple...
  - ... parfois sur des exemples faux ou fabriqués
- Des régularités dans l'acquisition du langage
  - Difficile après un certain âge (processus biologique, psychologique)
  - On peut apprendre n'importe quelle langue
  - +-Les mêmes étapes d'apprentissage quelque soit la langue

Analyse syntaxique : le problème

- Comment un locuteur sait-il qu'une phrase est bien formée ?
  - une tâche complexe
  - apprise par l'exemple...
  - ... parfois sur des exemples faux ou fabriqués
- Des régularités dans l'acquisition du langage
  - Difficile après un certain âge (processus biologique, psychologique)
  - On peut apprendre n'importe quelle langue
  - +-Les mêmes étapes d'apprentissage quelque soit la langue

  Diagnostiquer VS remédier, Incorrection VS incompréhension

→ Comment formaliser pour implémenter ?

Analyse syntaxique : les approches

Deux approches (comme toujours) :

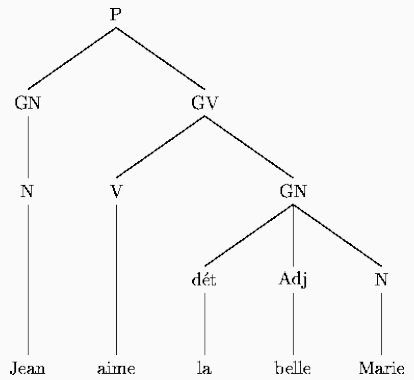
- Approches dites linguistiques
- Approches dites statistiques

Analyse syntaxique : approches linguistiques

Les constituants :
Certains segments possèdent une unité que l'on peut tester :
La phrase que la sœur de Robert a dite est une collection de mots ordonnés

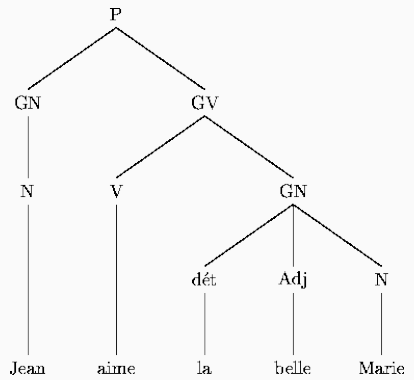
- Substitution
  - La phrase qu'**elle** a dite est une succession de mots ordonnés
- Mouvement
  - **C'est une succession de mots ordonnés**, la phrase que la sœur de Robert a dite.
- Question
  - Qu'a dit la sœur de Robert ? Une phrase.

Analyse syntaxique : un exemple



Source : Jérome Cardot - Université de Bretagne Occidentale

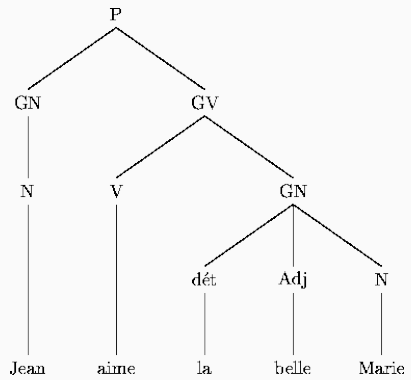
Analyse syntaxique : un exemple



- N, V, dét, Adj, N ⇒ **Symboles terminaux**

Source : Jérome Cardot - Université de Bretagne Occidentale

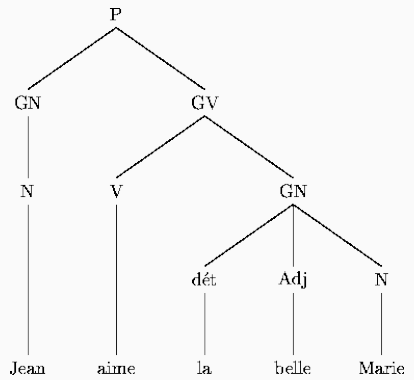
Analyse syntaxique : un exemple



- N, V, dét, Adj, N ⇒ **Symboles terminaux**
- GN, GV ⇒ **Variables**

Source : Jérome Cardot - Université de Bretagne Occidentale

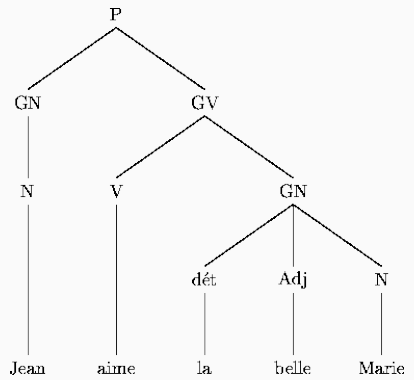
Analyse syntaxique : un exemple



- N, V, dét, Adj, N $\Rightarrow$ **Symboles terminaux**
- GN, GV $\Rightarrow$ **Variables**
- P $\Rightarrow$ **Axiome**

Source : Jérome Cardot - Université de Bretagne Occidentale

Analyse syntaxique : un exemple



- N, V, dét, Adj, N $\Rightarrow$ **Symboles terminaux**
- GN, GV $\Rightarrow$ **Variables**
- P $\Rightarrow$ **Axiome**
- $GV \rightarrow V + GN \Rightarrow$ **Règles de transformation**

Source : Jérome Cardot - Université de Bretagne Occidentale

Analyse syntaxique : formalisation

Une grammaire hors-contexte G = {V, Σ , R, S}

- V, ensemble fini de symboles non-terminaux
- S, axiome
- Σ, alphabet de symboles terminaux
- R, V x (V ∪ Σ) règles de transformation

Par exemple :
P → GN GV
GN → N
N → Det
GV → Ver GN
GN → Det Adj N

Analyse syntaxique : exercice

- Liste de terminaux : Det, Nom, Ver, Adj, Adv, Prep
- Liste des variables : GN, GV, GP
- Rédiger les règles de transformations qui reconnaissent ces phrases :

GN	GV	GP/GN
Le bateau	part	à la dérive
Les petits rochellais	aiment	les cagouilles
Trump	perd vraiment	

Analyse syntaxique : exercice

- Liste de terminaux : Det, Nom, Ver, Adj, Adv, Prep
- Liste des variables : GN, GV, GP
- Rédiger les règles de transformations qui reconnaissent ces phrases :

GN	GV	GP/GN
Le bateau	part	à la dérive
Les petits rochellais	aiment	les cagouilles
Trump	perd vraiment	

Les règles, c'est l'**axe syntaxique** : l'organisation de la phrase

Les terminaux c'est l'**axe paradigmatique** : par ex. on peut prendre n'importe quel verbe (ou presque) pour remplacer `perd`

Analyse syntaxique : exercice

- Liste de terminaux : Det, Nom, Ver, Adj, Adv, Prep
- Liste des variables : GN, GV, GP
- Rédiger les règles de transformations qui reconnaissent ces phrases :

GN	GV	GP/GN
Le bateau	part	à la dérive
Les petits rochellais	aiment	les cagouilles
Trump	perd vraiment	

Les règles, c'est l'**axe syntaxique** : l'organisation de la phrase

Les terminaux c'est l'**axe paradigmatique** : par ex. on peut prendre n'importe quel verbe (ou presque) pour remplacer `perd`

$$\textbf{GN} \rightarrow \text{Det} + \text{Nom} \mid \text{Det} + \text{Adj} + \text{Nom}$$

Analyse syntaxique : exercice

- Liste de terminaux : Det, Nom, Ver, Adj, Adv, Prep
- Liste des variables : GN, GV, GP
- Rédiger les règles de transformations qui reconnaissent ces phrases :

GN	GV	GP/GN
Le bateau	part	à la dérive
Les petits rochellais	aiment	les cagouilles
Trump	perd vraiment	

Les règles, c'est l'**axe syntaxique** : l'organisation de la phrase

Les terminaux c'est l'**axe paradigmatique** : par ex. on peut prendre n'importe quel verbe (ou presque) pour remplacer `perd`

GN $\rightarrow$ Det + Nom | Det + Adj + Nom
GV $\rightarrow$ Ver +GP | Ver + GN —Ver + Adv

Analyse syntaxique : exercice

- Liste de terminaux : Det, Nom, Ver, Adj, Adv, Prep
- Liste des variables : GN, GV, GP
- Rédiger les règles de transformations qui reconnaissent ces phrases :

GN	GV	GP/GN
Le bateau	part	à la dérive
Les petits rochellais	aiment	les cagouilles
Trump	perd vraiment	

Les règles, c'est l'**axe syntaxique** : l'organisation de la phrase

Les terminaux c'est l'**axe paradigmatique** : par ex. on peut prendre n'importe quel verbe (ou presque) pour remplacer `perd`

GN $\rightarrow$ Det + Nom | Det + Adj + Nom
GV $\rightarrow$ Ver +GP | Ver + GN —Ver + Adv
GP $\rightarrow$ Prep + GN

Analyse syntaxique : exercice

- Liste de terminaux : Det, Nom, Ver, Adj, Adv, Prep
- Liste des variables : GN, GV, GP
- Rédiger les règles de transformations qui reconnaissent ces phrases :

GN	GV	GP/GN
Le bateau	part	à la dérive
Les petits rochellais	aiment	les cagouilles
Trump	perd vraiment	

Les règles, c'est l'**axe syntaxique** : l'organisation de la phrase

Les terminaux c'est l'**axe paradigmatique** : par ex. on peut prendre n'importe quel verbe (ou presque) pour remplacer `perd`

GN $\rightarrow$ Det + Nom | Det + Adj + Nom
GV $\rightarrow$ Ver +GP | Ver + GN —Ver + Adv
GP $\rightarrow$ Prep + GN
P $\rightarrow$ GN +GV

Désambiguïsation

Désambiguïsation syntaxique

permis → permettre+V — permis+N :

- un mot = plusieurs analyses
- mots inconnus

→ Comment faire ? Qu'utilise l'humain ?

Désambiguïsation syntaxique

permis → permettre+V — permis+N :

- un mot = plusieurs analyses
- mots inconnus

→ Comment faire ? Qu'utilise l'humain ?
Il m'a permis d'habiter chez lui.

On va chercher à modéliser le contexte :

- Contexte thématique
- Contexte d'énonciation (spécialisation)

Désambiguïsation sémantique

≪ L'avocat est passé à table ≫

Quel(s) sens est(sont) activé(s) ?

Désambiguïsation sémantique

≪ L'avocat est passé à table ≫

Quel(s) sens est(sont) activé(s) ?

→ Le contexte, l'historique

Questions :

- Que garder en mémoire ?
- Comment faire sans historique et ≪ hors contexte ≫ ?

Algorithme de Lesk

Modélise le contexte de la phrase

- Fondé sur la sémantique latente (LSA)
- "Pour bien me connaître, apprends à connaître mes voisins"
- Comparer les différentes définitions et le voisinage en contexte
- Calculer l'intersection en "mots"

Algorithme de Lesk

Modélise le contexte de la phrase

- Fondé sur la sémantique latente (LSA)
- "Pour bien me connaître, apprends à connaître mes voisins"
- Comparer les différentes définitions et le voisinage en contexte
- Calculer l'intersection en "mots"
- Simple et efficace mais avec quelques limites...
  1. Ambiguïté des voisins
  2. Taille du voisinage (fenêtre)
  3. Taille des définitions

Algorithme de Lesk

Modélise le contexte de la phrase

- Fondé sur la sémantique latente (LSA)
- "Pour bien me connaître, apprends à connaître mes voisins"
- Comparer les différentes définitions et le voisinage en contexte
- Calculer l'intersection en "mots"
- Simple et efficace mais avec quelques limites...
  1. Ambiguïté des voisins
  2. Taille du voisinage (fenêtre)
  3. Taille des définitions
- Améliorations : élagage, apprentissage des voisinages, plusieurs définitions

Lesk : exercice

a- le orange est un peu flashy;
b- Il est passé à l'orange ;
c- Il a jeté la peau de l'orange;
d- il m'a passé l'orange
e- la rencontre entre les bleus et les oranges
f- le garage a fait une peau neuve à leur monoplace orange
g- La SA Orange France a sollicité l'implantation ...

1. **Nom Propre** Anciennement France Télécom, une entreprise française de télécommunications
2. **Nom** Fruit de l'oranger, arbre appartenant à la famille des rutacées ou agrumes, en général de couleur orange et de forme sphérique. L'orange est composée d'une écorce orange et contient des pépins, une chair juteuse, divisée de cloisons et de loges
3. **Nom** (Par ellipse) (Code routier) Feu orange
4. **Adj/Nom** Couleur tertiaire dans les deux systèmes (soustractif et additif), composée à partir du rouge et du jaune.

Lesk : exercice

a- le orange est un peu flashy;
b- Il est passé à l'orange ;
c- Il a jeté la peau de l'orange;
d- il m'a passé l'orange
e- la rencontre entre les bleus et les oranges
f- le garage a fait une peau neuve à leur monoplace orange
g- La SA Orange France a sollicité l'implantation ...

1. **Nom Propre** Anciennement France Télécom, une entreprise française de télécommunications
2. **Nom** Fruit de l'oranger, arbre appartenant à la famille des rutacées ou agrumes, en général de couleur orange et de forme sphérique. L'orange est composée d'une écorce orange et contient des pépins, une chair juteuse, divisée de cloisons et de loges
3. **Nom** (Par ellipse) (Code routier) Feu orange
4. **Adj/Nom** Couleur tertiaire dans les deux systèmes (soustractif et additif), composée à partir du rouge et du jaune.

1. (g) ; 2. (c, d ... voire b !) ; 3. (b) ; 4. (a, e, f)

"Extension" de LSA : Word Embeddings

- D'une approche "descriptive" à une approche prédictive :
- $\rightarrow$ prédire les contextes en fonction des mots (et inversement)

"Extension" de LSA : Word Embeddings

- D'une approche "descriptive" à une approche prédictive :
- $\rightarrow$ prédire les contextes en fonction des mots (et inversement)

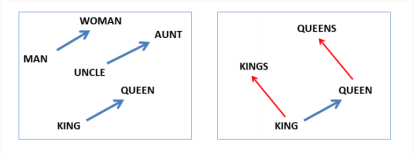


Figure 3: Mikolov *et al.* 2013 : : analogies sémantiques

"Extension" de LSA : Word Embeddings

- D'une approche "descriptive" à une approche prédictive :
- $\rightarrow$ prédire les contextes en fonction des mots (et inversement)

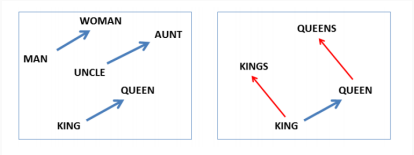


Figure 3: Mikolov *et al.* 2013 : : analogies sémantiques

Figure 4: Hamilton, Leskovec, Jurafsky : néologie sémantique