Statistiques pour la linguistique

Damien Nouvel

Plan

1. Prétraitements

2. Annotation morpho-syntaxique

3. Statistiques textuelles

Données du corpus

- Corpus comme ensemble de **documents** (ou parties)
- Deux éléments
  - Le **contenu** (diverses structures)
  - Les **métadonnées** :
    - Auteur
    - Date de création
    - Mots-clés
    - ...

  ⇒ Le titre, le résumé sont-ils du contenu ou des métadonnées ?
- Focale sur le **contenu** comme **texte brut** :
  - Suite de caractères **UTF-8** segmentés en mots
  - Peu de prise en compte de la mise en page

⇒ Un document, un fichier `doc1.txt`

⇒ Prétraitements pour accéder à la linguistique

⇒ Suite/ensemble de **tokens** porteurs de **sens**

Segmentation

- Séparer un texte (en phrases puis) en mots, les **tokens**
- Difficultés pour le français :
  - **Clitiques**, composition semi-soudées
  - **Expressions polylexicales**

  ⇒ Utilisation d'automates

⇒ Utilisation répandue de **TreeTagger**

⇒ En python, dans des librairies (NLTK, Spacy, etc.)

Représentation séquentielle

- Notations
  - **Séquence** : $\langle c_1, c_2 \dots c_n \rangle$
  - **Item** : c_1 élément de la séquence

$\Rightarrow$ Suite de lettres, de mots, de caractères

- Exemple *"Il m'a parlé du porte-avions Charles de Gaulle"*

Il	m'a	parlé	du	porte-	avions	Charles	de	Gaulle

Il	m'	a	parlé	du	porte-	avions	Charles	de	Gaulle

Il	m'	a	parlé	du	porte-avions	Charles de Gaulle

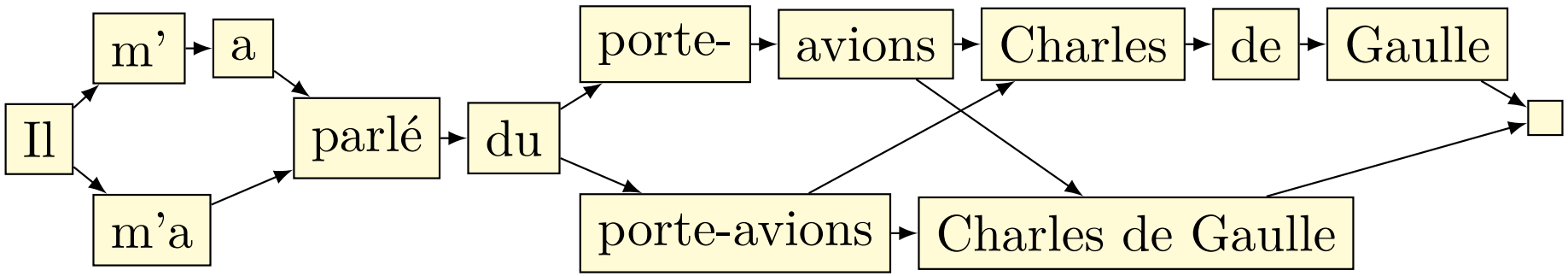
Ambiguïtés et graphes

- Formalisme pour les ambiguïtés :
  - **Nœud** : mot ou suite de mots
  - **Arc** (flèche) : choix d'un chemin

⇒ Chemin est une analyse possible

⇒ Combinatoire des analyses

- Exemple *"Il m'a parlé du porte-avions Charles de Gaulle"*



Autres représentations

- De nombreuses autres représentations possibles :
  - Arbres syntaxiques (constituants)
  - Graphes de dépendances
  - Sacs de mots
  - Chaînes de coréférence
  - “Cadres” sémantiques (frames)
  - ...

$\Rightarrow$ Et pour chacune, N possibilités pour faire des statistiques ...

Plan

1. Prétraitements

2. Annotation morpho-syntaxique

3. Statistiques textuelles

La catégorisation morpho-syntaxique

- Affecter des **catégories morpho-syntaxiques** aux **tokens**
- Un choix toujours ambigü
  - Selon le **lexique**
  - Selon le corpus **d'entraînement**
  - Selon l' **approche** utilisée (vote, HMM, N-grammes, CRF)

  $\Rightarrow$ Prise de décision selon les mots et le contexte

$\Rightarrow$ Pour une phrase de n tokens $< m_1 \dots m_n >$, déterminer les étiquettes associées $< e_1 \dots e_n >$ qui sont les plus **vraisemblables**

$\Rightarrow$ $max_{<e_1 \dots e_n>} P(< e_1 \dots e_n > | < m_1 \dots m_n >)$

- Corpus étiqueté (TreeTagger) au format "Brown corpus" :

```
Passepartout/NAM demeura/VER seul/ADJ dans/PRP
la/DET maison/NOM de/PRP Saville-row/NAM ./SENT
```

Etiquetage par classes majoritaires

- Hypothèse d'**indépendance** sur les mots et les étiquettes :

$\Rightarrow P(< e_1 \dots e_n > | < m_1 \dots m_n >) = \prod_i P(e_i|m_i)$

- Statistiques simples :
  - Fréquences des **mots** $F(m)$
  - Fréquences des **étiquettes** $F(e)$
  - Fréquence des **associations mots-étiquettes** $F(m, e)$
- Étiquette qui **maximise la probabilité** sachant le mot :
  - Pour un mot donné, $P(e|m) = \frac{P(e,m)}{P(m)}$
  - Or (corpus de taille N), $P(e, m) = \frac{F(m,e)}{N}$ et $P(m) = \frac{F(m)}{N}$
  - Comparaisons pour un m donné : $F(m)$ n'a pas d'influence

  $\Rightarrow$ Pour un mot, **étiquette "majoritaire"** : $max_e F(m, e)$

```
me = {'avoir': {'VER': 30, 'NOM': 5}, 'auras': {'VER': 17}}
e = sorted(me['avoir'].items(), key=lambda x: x[1])[-1][0]
```

Modèle de Markov Caché

⇒ Quelle suite d'états a pu "générer" la phrase mot à mot ?

- Décomposition de la probabilité :
  - Approche **bayésienne** :
  - ⇒ $P(\langle e_1 \dots e_n \rangle | \langle m_1 \dots m_n \rangle) = \frac{P(\langle (e_1, m_1) \dots (e_n, m_n) \rangle)}{P(\langle m_1 \dots m_n \rangle)}$
  - Hypothèse **markovienne** de contexte limité :
  - ⇒ $P(< (e_1, m_1) \dots (e_n, m_n) >= P(e_1, m_1) * \prod_i P(e_i, m_i | e_{i-1})$
  - Vraisemblance selon les **générations** et **transitions** :
  - ⇒ $P(e_i, m_i | e_{i-1}) \sim P(e_i | e_{i-1}) * P(m_i | e_i)$
- Ajout des statistiques :
  - D'**émission** des mots : $P(m|e) = \frac{F(m,e)}{F(e)}$
  - De **transition** d'étiquettes (bigrammes : $P(e_1|e_2) = \frac{F(e_1,e_2)}{F(e_2)}$
- Suite d'étiquettes qui maximise la probabilité de génération :

⇒ $max_{<e_1 \dots e_n>} P(m_1|e_1) * \prod_{i=1\dots n} P(e_i|e_{i-1}) * P(m_i|e_i)$

Utilisation des lexiques

- Objectifs multiples :
  - **Catégoriser** les mots (morphologie, syntaxe, etc.)
  - Affecter des **classes sémantiques** aux **tokens**
  - Constituer ou utiliser une **terminologie**
  - **Normalisation** de termes ou d'entités spécifiques

⇒ Inventaire de **mots** ou d'**expressions** et de propriétés

⇒ Reconnaissance par **automates déterministes**

- Exemple de difficultés rencontrées avec les lexiques :
  - **Synonymie** : plusieurs mots pour la même sémantique

  ⇒ Agrandit la taille du lexique
  - **Homonymie** : un même mot (typographique ou phonétique) peut avoir de multiple sens

  ⇒ Ambiguïté du mot
  - **Métonymie** : la sémantique d'un mot change en contexte

  ⇒ Difficulté de prévoir le phénomène

Plan

1. Prétraitements

2. Annotation morpho-syntaxique

3. Statistiques textuelles

Représentation matricielle

- Hypothèses
  - Corpus séparé en **documents** ou **parties**
  - Textes déjà segmenté (tokenisé)
  - Corpus comme matrice termes / documents (sacs de mots)

⇒ **Fréquences** des termes dans les documents

	t_1	t_2	t_3	...
d_1	f_{11}	f_{12}	f_{13}	...
d_2	f_{21}	f_{22}	f_{23}	...
d_3	f_{31}	f_{32}	f_{33}	...
...	...	...	...	...

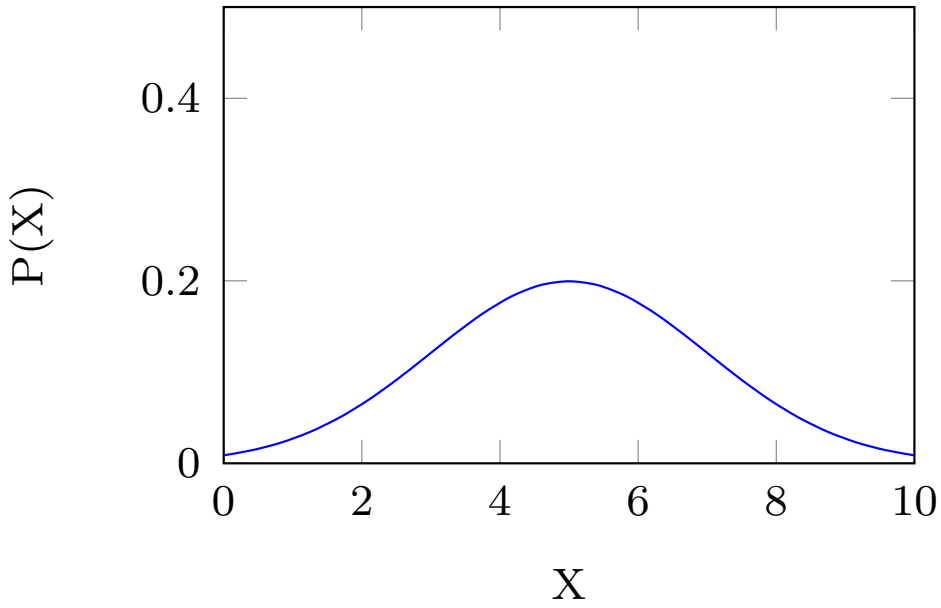
⇒ *Vector Space Model*

- Calculs statistiques facilités
  - Taille moyenne des documents
  - Fréquence moyenne d'un terme par document
  - Coocurrences de termes

Loi normale

- Principes généraux
  - Aucun *apriori* sur la répartition des données
  - Paramètres : moyenne, écart-type

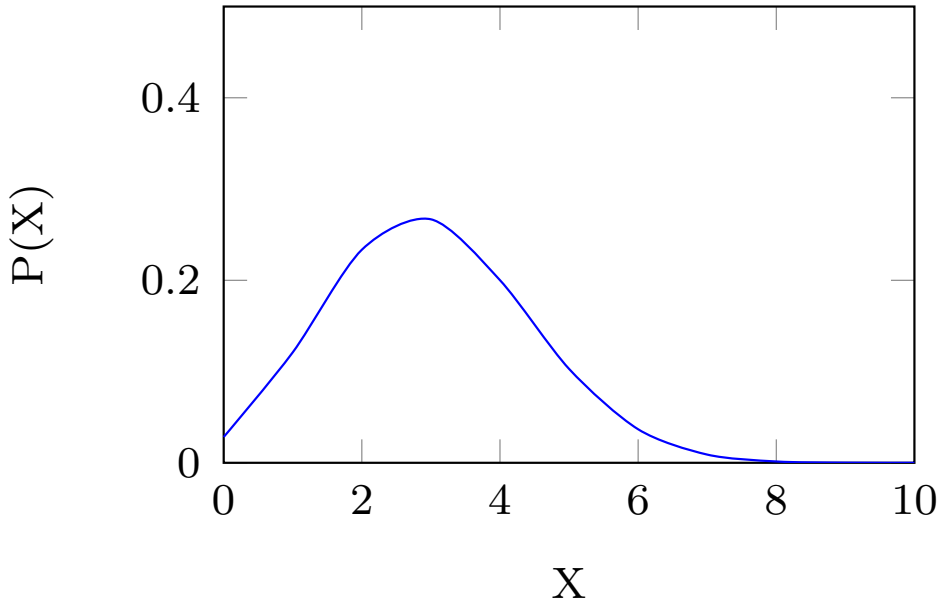
  $\Rightarrow$ Peu adaptée aux fréquences de termes
- Formule : $P(X = x) = \frac{1}{\sigma_x \sqrt{2\pi}} exp\left(-\frac{(x - \mu_x)^2}{2\sigma_x^2}\right)$
- Courbe ($\mu_x = 5$, $\sigma_x = 2$) :



Loi binomiale

- Principes généraux
  - **Répétition d'une épreuve** n fois avec remise
  - Combien de « succès » ?

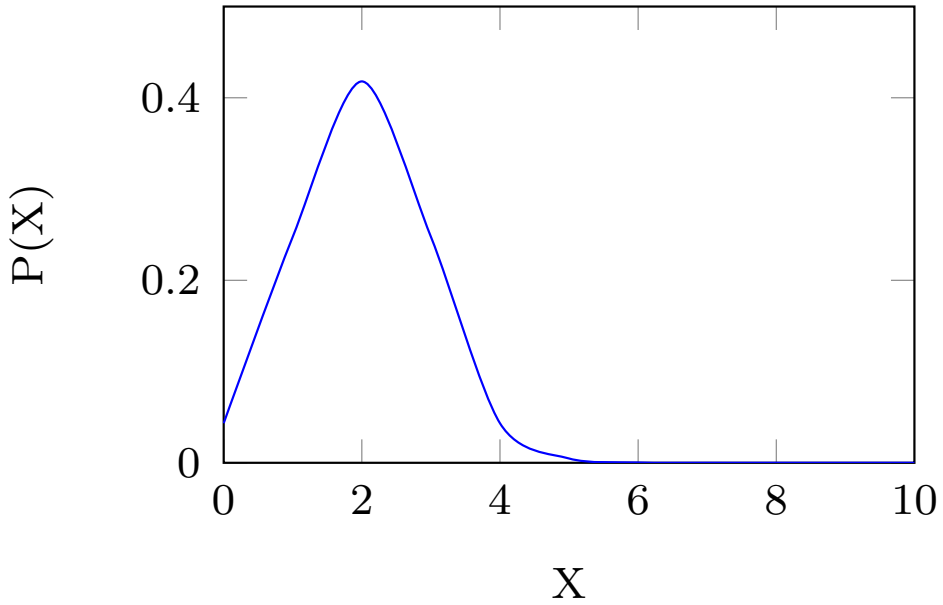
  $\Rightarrow$ Entre 0 et n, selon la probabilité
  - Paramètres : probabilité p, répétition n
- Formule : $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Courbe ($p = 0,3$, $n = 10$) :



Loi hypergéométrique

- Principes généraux
  - Répétition d'une épreuve n fois **sans remise**
  - Combien de « succès » ?

  $\Rightarrow$ Entre 0 et n, avec une probabilité décroissante
  - Paramètres : probabilité p, répétition n, nombre total N
- Formule : $P(X = k) = \binom{pN}{k} * \binom{(1-p)N}{n-k} / \binom{N}{n}$
- Courbe ($p = 0,3$, $n = 10$, $N = 20$) :



Calcul des spécificités

- Statistique sur les **fréquences des termes par partie**
  - Formule sur la matrice termes / documents (parties)
    - f_{ij} (fréquence dans une partie i d'un terme j)
    - $T = \sum_{ij} fij$ (taille totale du corpus)
    - $d_i = \sum_j f_{ij}$ (taille d'une partie i)
    - $t_j = \sum_i f_{ij}$ (nombre total d'occurence du terme j

$$\Rightarrow P(f_{ij} = k) = \frac{\binom{t_j}{k} * \binom{T-t_j}{d_i-k}}{\binom{T}{d_i}}$$

- **Spécificités** pour la partie au regard du corpus
  - Probabilité faible (fréquence inattendue) par **seuil** (0,05)
  - Spécificités **positives** : fréquence forte, sur-représentation
  - Spécificités **négatives** : fréquence faible, sous-représentation

  $\Rightarrow$ Permet de **caractériser** la sous-partie du corpus
- On peut travailler sur plusieurs **partitions** du corpus