

Introduction

Damien Nouvel



Plan

1. Généralités
2. Contexte général
3. Corpus et statistiques
4. Visées applicatives

Séances et modalités de contrôle

▸ Séances

- 12 séances
- Chaque séance : **cours** et **exercices** sur machine
- Cours S2 : *apprentissage automatique*

▸ Modalités de contrôle

- Un **examen** final (50%)
- Un **contrôle** (50%)
- Une note de **contrôle continu** (+/-1 point)

Contenu du cours (prévisionnel)

▸ Progression

- Introduction
- Programmer en Python
- Éléments en théorie des probabilités et en statistiques
- Prétraitements de corpus
- Théorie de l'information et mesures d'entropie
- Statistiques pour l'évaluation
- Classifieurs
- Projet

Plan

1. Généralités
2. Contexte général
3. Corpus et statistiques
4. Visées applicatives

Informatique et langage



- ▶ **Grace Hopper** : A0, premier langage compilé (1951)

Traitement Automatique des Langues (TAL)

- ▶ Le langage
 - **Expressivité** (implicite, déductions)
 - **Générativité** / compositionnalité
 - **Modes d'expression** : oral, écrit, signes, etc.
 - **Evolution** par conventions sociales

 - ▶ L'informatique
 - **Calcul** binaire, entiers, flottants, etc.
 - **Procédures** par séquence d'instructions
 - **Copies exactes**
 - **Télécommunications** et essor des nouvelles technologies
- ⇒ **Discipline** issue de l'essor de l'utilisation de l'informatique pour que des **humains** manipulent le **langage** par d'autres moyens et à une **autre échelle**

Plan

1. Généralités
2. Contexte général
3. Corpus et statistiques
4. Visées applicatives

Un corpus : quelles données, pour quoi faire

⇒ Ensemble de documents (textuels) ...

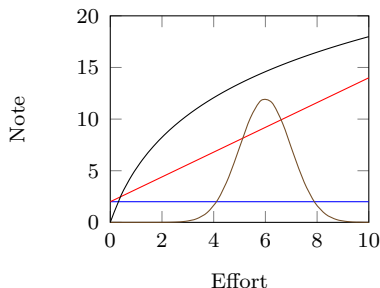
- ▶ Des *classifications* associées, par
 - auteur(s) et éditeur
 - langue
 - date de publication
 - style (roman, journal, publication, administratif, etc.)
 - thèmes (science-fiction, politique, sport, technique, etc.)

⇒ Création d'**index** pour trouver les documents

- ▶ **Ressources** pour
 - linguistique
 - publications scientifiques (par ex. biomédical)
 - sciences de l'homme (sociologie)
 - traitement du langage parlé
 - etc.

⇒ Données disponibles pour **traitements** et études

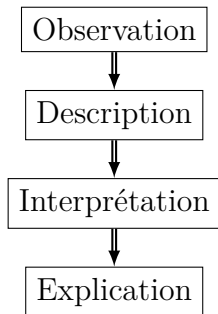
Les outils statistiques



- ▶ Les outils statistiques vont aider à modéliser des **corrélations** entre des données selon des **paramètres à déterminer**
- ▶ Quelques notions de base
 - **Fréquence** : nombre d'occurrences
 - **Probabilité** : nombre compris dans $[0, 1]$

Statistiques pour les sciences

► Découvertes scientifiques



► **Corrélation** n'est pas **causalité**

- Parler russe et boire de la vodka
- Travailler dans un garage et devenir millionnaire

Épistémologie : règles logiques

- ▶ Expliquer un phénomène selon des **conditions**
 - **Nécessaire** : $\neg a \Rightarrow \neg b$
 - **Suffisante** : $a \Rightarrow b$
 - **Nécessaire et suffisante** : $a \Leftrightarrow b$
- ▶ **Règles logiques** pour l'inférence
 - **Dédution** :

$$\frac{(a \Rightarrow b) \wedge a}{b}$$

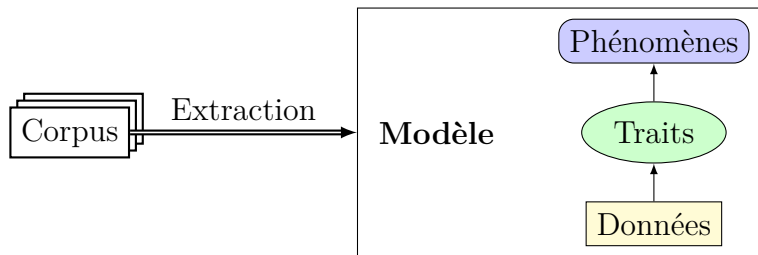
- **Abduction** :

$$\frac{(a \Rightarrow b) \wedge b}{a}$$

- **Induction** :

$$\frac{a \wedge b}{a \Rightarrow b}$$

Corpus, modèle, tâche



- ▶ Les corpus permettent de **décrire des phénomènes**
 - **Paramétrage** du modèle (plus ou moins *supervisé*)
 - **Evaluation** du modèle sur d'autres données
 - **Utilisation** du modèle pour d'autres tâches
- ⇒ Les **statistiques** permettent de **concevoir**, de **paramétrer** et d'**évaluer** l'adéquation du modèle pour une tâche donnée

Plan

1. Généralités
2. Contexte général
3. Corpus et statistiques
4. Visées applicatives

Exemples d'applications

- ▶ Niveaux de **représentation** du langage
 - Acoustique (parlé)
 - Phonétique (parlé)
 - Morphologie
 - Syntaxe
 - Sémantique
 - Pragmatique

- ▶ **Statistiques de corpus** pour
 - **Indexation** de documents et **recherche d'information**
 - Correction **orthographique** et **grammaticale**
 - **Reconnaissance** automatique de l'**écriture**
 - **Reconnaissance** automatique de la **parole**
 - **Traduction automatique**
 - ... (et bien d'autres)

Pourquoi les statistiques en TAL

- ▶ **Historique** rapide et incomplet des **approches TAL**
 - **Symboliques** : automates, transducteurs
 - **Heuristiques** : logique, affectation manuel de poids de règles
 - **Numériques**, dont par exemple
 - Réseaux de neurones
 - Maximum d'entropie / régression logistique
 - Arbres de décision
 - Approches bayésiennes
 - Chaînes de Markov
 - SVM, CRF
- ⇒ Plus de **numérique**, plus de **statistiques**
- ▶ Importance accrue des statistiques pour la linguistique
 - Mise à disposition d' **importants volumes** de textes
 - Exigence de **robustesse** des applications

Littérature

- ▶ Quelques références
 - Sites web
 - Michèle Jardino <http://archives.limsi.fr/Individu/jardino/coursTCAN2005.pdf>
 - Marti Hearst <http://courses.ischool.berkeley.edu/i256/f06/sched.html>
 - Christopher Manning <http://web.stanford.edu/class/cs276b/>
 - Peter Norvig <http://norvig.com/chomsky.html>
 - Revue TAL <http://atala.org/~Revue-TAL->
 - Revue CL <http://cljournal.org/>
 - Livres
 - Initiation aux méthodes de la stat. linguistique (Muller, 1993)
 - Modèles statistique pour l'accès à l'information textuelle (Gaussier, Yvon, 2011)
 - Statistique textuelle (Lebart, Salem, 1994)
 - Foundation of Statistical NLP (Manning, Schütze, 1999)
 - The Handbook of CL and NLP (Clark, Fox, Lappin, 2010)
 - Natural Language Annot. for Machine Learning (Pustejovsky, Stubbs, 2012)