Le traitement automatique des langues (TAL) et les statistiques : panorama et problématiques

Cyril Grouin Pierre Zweigenbaum

Rendez-vous SFdS ≪ *Méthodes et Logiciels* ≫

11 avril 2013

Traitement automatique des langues

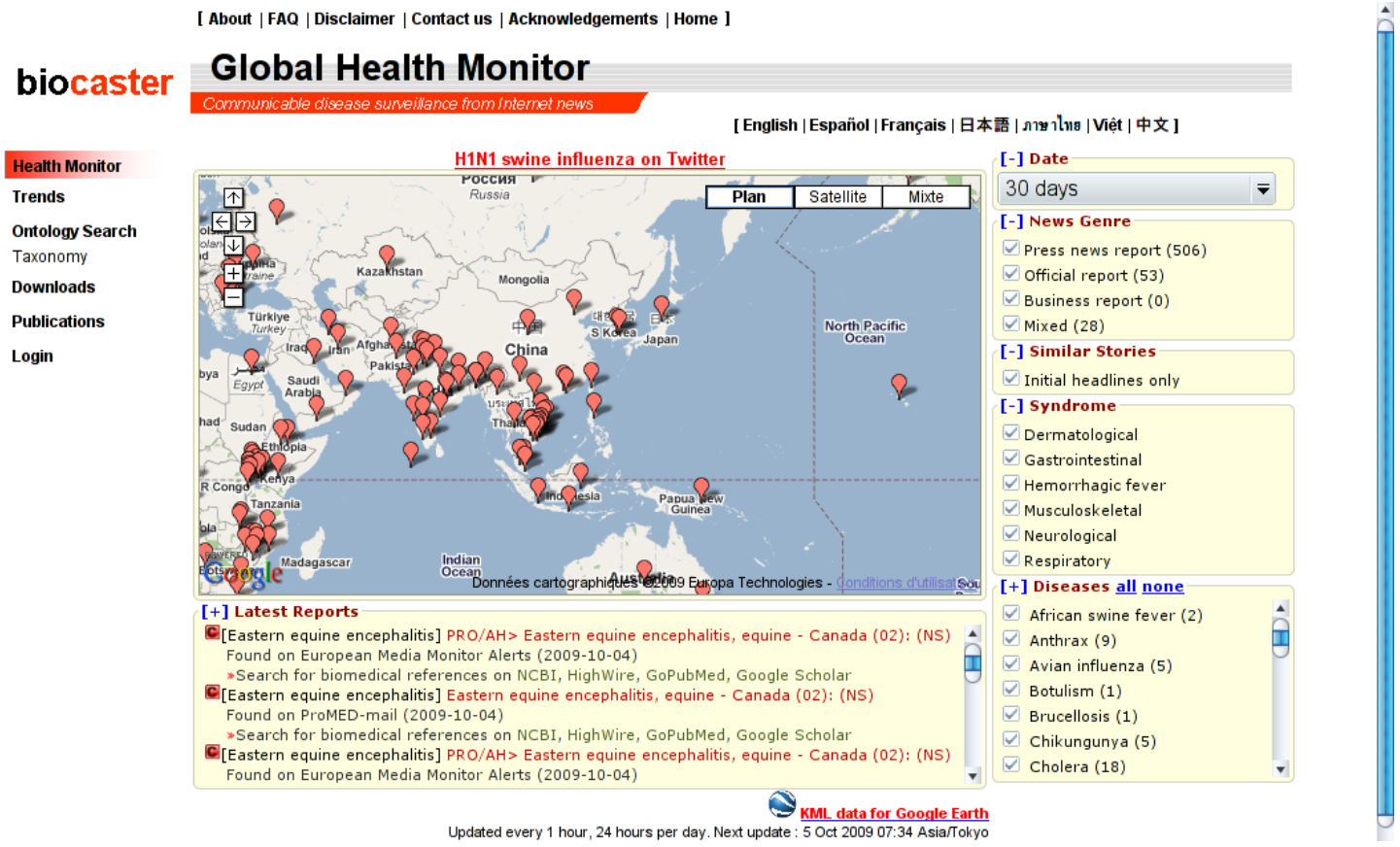
Traitement Automatique des Langues

- **Définition :** « *NLP (...) is the study of the computational, mathematical and statistical properties of natural languages and systems for processing languages* » [Gazdar, 1996] ;
- **Résumé :** ensemble des processus informatiques mis en œuvre pour traiter des éléments formulés en langue naturelle :
  - analyse → accéder au sens ;
  - production → retranscrire du sens.

Exemple : Extraction d'information, nouvelles et épidémies

BioCaster

http://biocaster.nii.ac.jp/



Traitement Automatique des Langues

Objectif : tenter de formaliser les langues naturelles

Exemple :

- **Grammaire :** ensemble des règles qui décrivent le fonctionnement d'une langue
  - Formation du pluriel des noms : ajout d'un « s »
  - Formation d'un syntagme nominal :

$$\text{SN} \longrightarrow \text{Det Nom}$$
$$\text{SN} \longrightarrow \text{Det Nom Adj}$$

- Exceptions, variantes
  - mots invariables : *une souris/des souris* ;
  - désinences différentes : *un cheval/des chevaux* ;
  - formes différentes : *un œil/des yeux*.
  - nombreuses variantes de construction :

$$\text{SN} \longrightarrow \text{Det Adj Nom}$$

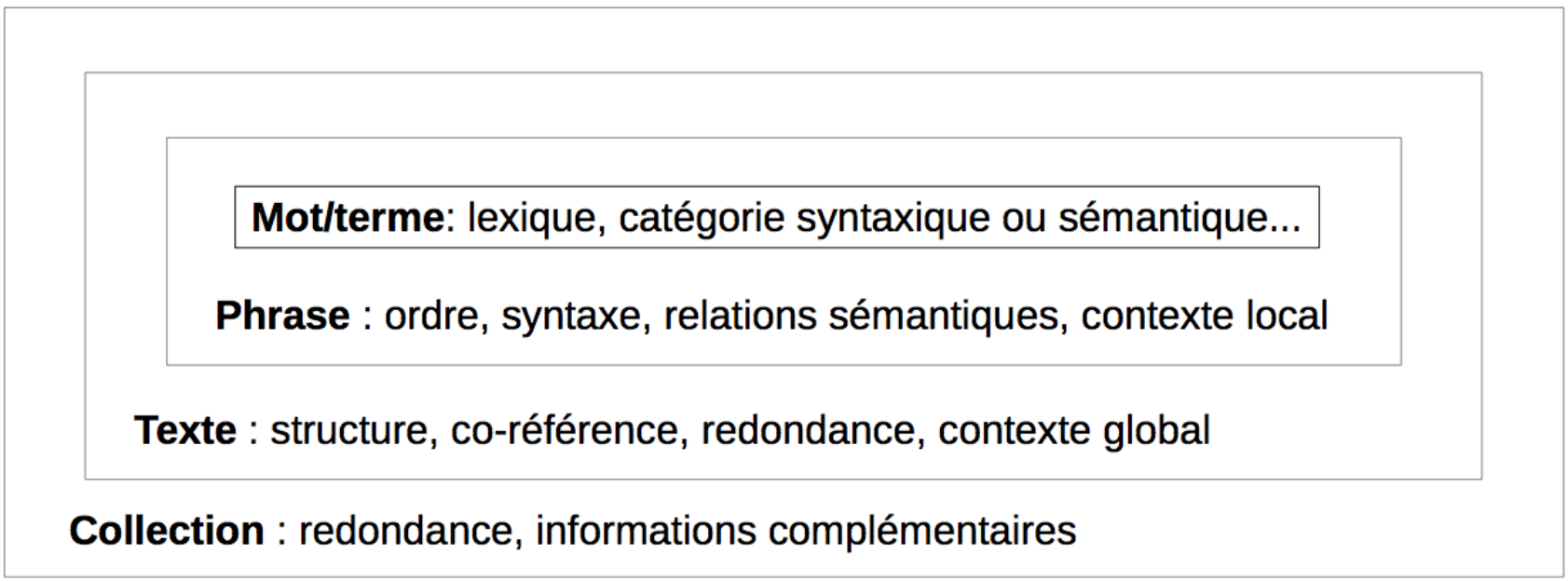
Traitement Automatique des Langues

Sources de variation

- **modalités d'expression** → techniques de traitement :
  - écrit *(presse écrite, documents cliniques)* ;
  - oral : monologues *(presse radio et télé)* vs. interactions *(débats, question/réponse)* ;
  - signé : discours en langue des signes (LSF).
- **types de langue** → vocabulaire et structures syntaxiques :
  - générale *(quotidienne)* ;
  - spécialité *(médecine, politique)*.
- **origine des données** → prétraitements :
  - données propres et de qualité ;
  - données bruitées *(OCR, téléphone, forums internet)*.

Traitement Automatique des Langues

- **niveaux d'intervention**



Source : Brigitte Grau

Traitement Automatique des Langues

Exemples d'applications

- production de texte ;
- lecture et compréhension de texte ;
- traduction ;
- accès à l'information ;
- systèmes interactifs ;
- indexation automatique ;
- etc.

Source : Diana Santos, `http://www.portugues.mct.pt/Diana/bitTutEval.html`

Traitement Automatique des Langues

Production de texte

- correction orthographique, grammaticale, stylistique ;
- vérification de langage contrôlé ;
- production de documentation automatique ;
- aide à la numérisation *(correction d'OCR)* ;
- etc.

Source : Diana Santos, `http://www.portugues.mct.pt/Diana/bitTutEval.html`

Traitement Automatique des Langues

Lecture et compréhension

- extraction d'information *(repérage d'entités nommées)* ;
- résumé automatique ;
- fouille de texte ;
- recherche d'information multilingue ;
- etc.

Source : Diana Santos, `http://www.portugues.mct.pt/Diana/bitTutEval.html`

Traitement Automatique des Langues

Traduction

- traduction automatique ;
- aide à la traduction ;
- traduction de la parole ;
- transcription automatique en temps réel ;
- etc.

Source : Diana Santos, `http://www.portugues.mct.pt/Diana/bitTutEval.html`

Traitement Automatique des Langues

Accès à l'information

- systèmes d'information par téléphone ;
- systèmes question/réponse ;
- recherche sur internet ;
- etc.

Source : Diana Santos, `http://www.portugues.mct.pt/Diana/bitTutEval.html`

Traitement Automatique des Langues

Systèmes interactifs

- commandes vocales ;
- interaction avec des assistants automatiques ;
- rapports automatiques ;
- jeux interactifs ;
- etc.

Source : Diana Santos, `http://www.portugues.mct.pt/Diana/bitTutEval.html`

Traitement Automatique des Langues

Indexation

- création semi-automatique de thesaurus ;
- indexation de documents ;
- catalogue d'images ;
- création semi-automatique de terminologie bilingue ;
- etc.

Source : Diana Santos, `http://www.portugues.mct.pt/Diana/bitTutEval.html`

Traitement Automatique des Langues

Autres applications

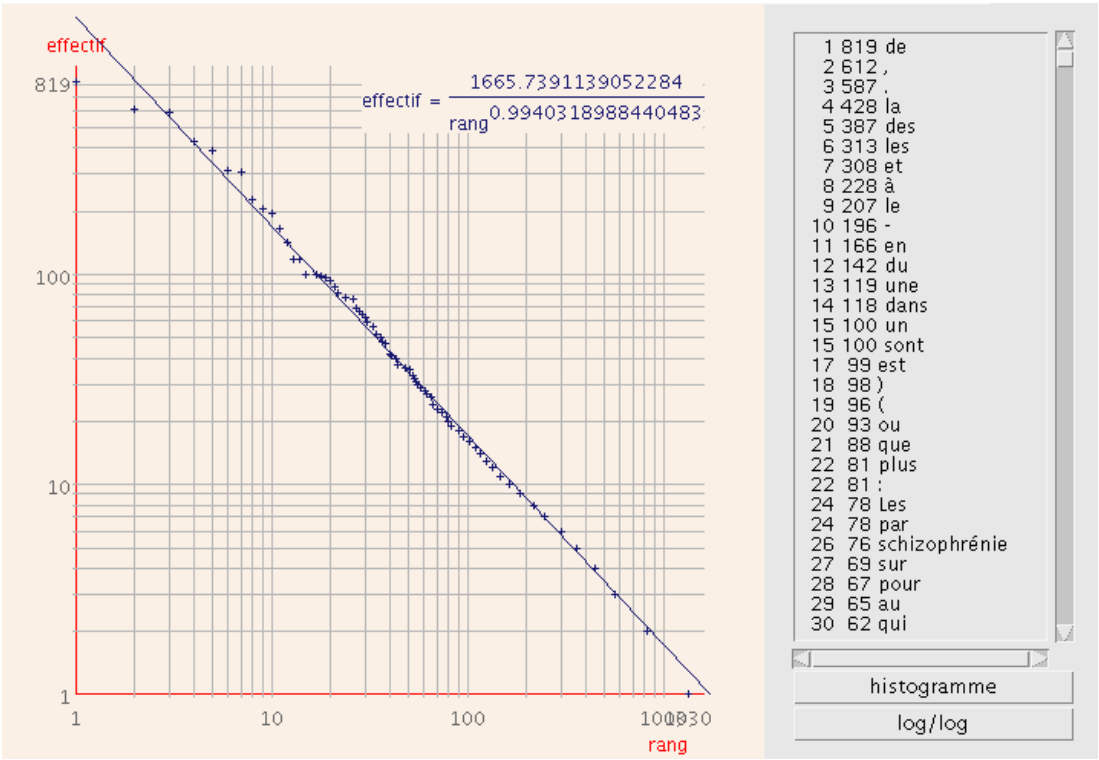
- études sociologiques ;
- analyse du vocabulaire en domaine de spécialité ;
- études linguistiques en diachronie ;
- etc.

Source : Diana Santos, `http://www.portugues.mct.pt/Diana/bitTutEval.html`

Omniprésence des modèles statistiques en TAL

La loi de Zipf-Mandelbrot

- le produit de la fréquence f de n'importe quel mot par son rang r est (approximativement) constant
- $f * r = C$
- grand nombre d'événements (mots) rares



- `http://users.info.unicaen.fr/~giguet/java/zipf.html`

Connaissances et décisions certaines ou probables

L'ambiguïté est partout

- Catégorie syntaxique d'un mot :
  - *est* : V, N, A
- Traduction d'un mot :
  - *tableau* (fr) → (en) table, blackboard
- Rattachement d'un syntagme prépositionnel :

  V SN de SN → rattachement au V
  V SN de SN → rattachement au SN

Connaissances et décisions certaines ou probables

L'ambiguïté est partout

- Catégorie syntaxique d'un mot :
  - *est* : (0.90) V, (0.05) N, (0.05) A
- Traduction d'un mot :
  - *tableau* (fr) → (en) (0.80) table, (0,20) blackboard
- Rattachement d'un syntagme prépositionnel :
  - (0.20) V SN de SN → rattachement au V
  - (0.80) V SN de SN → rattachement au SN

On peut chercher à la probabiliser.

Modèles statistiques en TAL

Applications du TAL

- **moteurs de recherche :** *100% of major players are trained and probabilistic. Their operation cannot be described by a simple function.*
- **reconnaissance de la parole :** *100% of major systems are trained and probabilistic, mostly relying on probabilistic hidden Markov models.*

Source : Peter Norvig, `http://norvig.com/chomsky.html`

Modèles statistiques en TAL

Applications du TAL

- **traduction automatique :** *100% of top competitors in competitions such as NIST use statistical methods. [...] Of the 4000 language pairs covered by machine translation systems, a statistical system is by far the best for every pair.*
- **systèmes question/réponse :** *this application is less well-developed, and many systems build heavily on the statistical and probabilistic approach used by search engines.*

Source : Peter Norvig, `http://norvig.com/chomsky.html`

Modèles statistiques en TAL

Composants du TAL

- **désambiguïsation sémantique :** *100% of top competitors at the SemEval-2 competition used statistical techniques ; most are probabilistic ; some use a hybrid approach incorporating rules from sources such as Wordnet.*
- **résolution des coréférences :** *The majority of current systems are statistical, although we should mention the system of Haghighi and Klein, which can be described as a hybrid system that is mostly rule-based rather than trained, and performs on par with top statistical systems.*

Source : Peter Norvig, `http://norvig.com/chomsky.html`

Modèles statistiques en TAL

Composants du TAL (suite)

- **étiquetage en parties du discours :** *Most current systems are statistical. The Brill tagger stands out as a successful hybrid system : it learns a set of deterministic rules from statistical data.*
- **analyse syntaxique :** *There are many parsing systems, using multiple approaches. Almost all of the most successful are statistical, and the majority are probabilistic (with a substantial minority of deterministic parsers).*

Source : Peter Norvig, `http://norvig.com/chomsky.html`

Repérage d'entités nommées

Repérage d'entités nommées (REN)

- **définition :** éléments du texte catégorisables sur le plan sémantique.
  - 1992 [Coates-Stephens, 1992] : noms de personnes ;
  - 1995, challenge MUC-6 [Grishman and Sundheim, 1996] :
    - noms de personnes : *Victor Hugo* ;
    - noms de lieux : *Europe, Versailles* ;
    - noms d'organisations : *OCDE, BNPParibas* ;
    - données numériques : dates, heures et montants.
  - Depuis :
    - sous-spécifications *(lieux → villes)* [Fleischman, 2001],
    - et nouvelles catégories *(fonctions, produits)* [Sekine, 2004].
- **objectif :** répondre à des questions de base :
  → *Qui ? Quoi ? Où ? Quand ? Comment ?*

Repérage d'entités nommées (REN)

Exemple :

Le sculpteur César est mort hier à Paris , à l' âge de 77 ans . Comme Yves Montand , son ami , César Baldaccini , fils d' émigrés italiens , naît en 1921 dans un quartier très pauvre de Marseille . Le sculpteur , fou de Brancusi et de Picasso , se répète beaucoup .

Repérage d'entités nommées (REN)

Exemple : noms de personnes

Le sculpteur [personne César] est mort hier à Paris , à l' âge de 77 ans . Comme [personne Yves Montand] , son ami , [personne César Baldaccini] , fils d' émigrés italiens , naît en 1921 dans un quartier très pauvre de Marseille . Le sculpteur , fou de [personne Brancusi] et de [personne Picasso] , se répète beaucoup .

Repérage d'entités nommées (REN)

Exemple : noms de lieux

Le sculpteur [personne César] est mort hier à [lieu Paris], à l'âge de 77 ans. Comme [personne Yves Montand], son ami, [personne César Baldaccini], fils d'émigrés italiens, naît en 1921 dans un quartier très pauvre de [lieu Marseille]. Le sculpteur, fou de [personne Brancusi] et de [personne Picasso], se répète beaucoup.

Repérage d'entités nommées (REN)

Exemple : noms de fonctions

Le [fonction sculpteur] [personne César] est mort hier à [lieu Paris] , à l' âge de 77 ans . Comme [personne Yves Montand] , son ami , [personne César Baldaccini] , fils d' émigrés italiens , naît en 1921 dans un quartier très pauvre de [lieu Marseille] . Le [fonction sculpteur] , fou de [personne Brancusi] et de [personne Picasso] , se répète beaucoup .

Repérage d'entités nommées (REN)

Exemple : dates, données numériques

Le [fonction sculpteur] [personne César] est mort [date hier] à [lieu Paris] , à l' âge de [âge 77 ans] . Comme [personne Yves Montand] , son ami , [personne César Baldaccini] , fils d' émigrés italiens , naît en [date 1921] dans un quartier très pauvre de [lieu Marseille] . Le [fonction sculpteur] , fou de [personne Brancusi] et de [personne Picasso] , se répète beaucoup .

Apprentissage (statistique) supervisé

Apprentissage supervisé

- ≪ *l'apprentissage artificiel a pour objectif la mise au point de programmes capables d'apprendre à partir de leur expérience, c'est-à-dire de changer leur structure interne ou la valeur de leurs paramètres en fonction de leur expérience de manière à améliorer leurs performances futures.* ≫ [Wisniewski, 2007]
- ≪ *plutôt que d'écrire une spécification formelle du comportement du programme, le programmeur fournit une base d'apprentissage composée d'exemples d'entrée et leur sortie attendue* ≫ [ibid.]

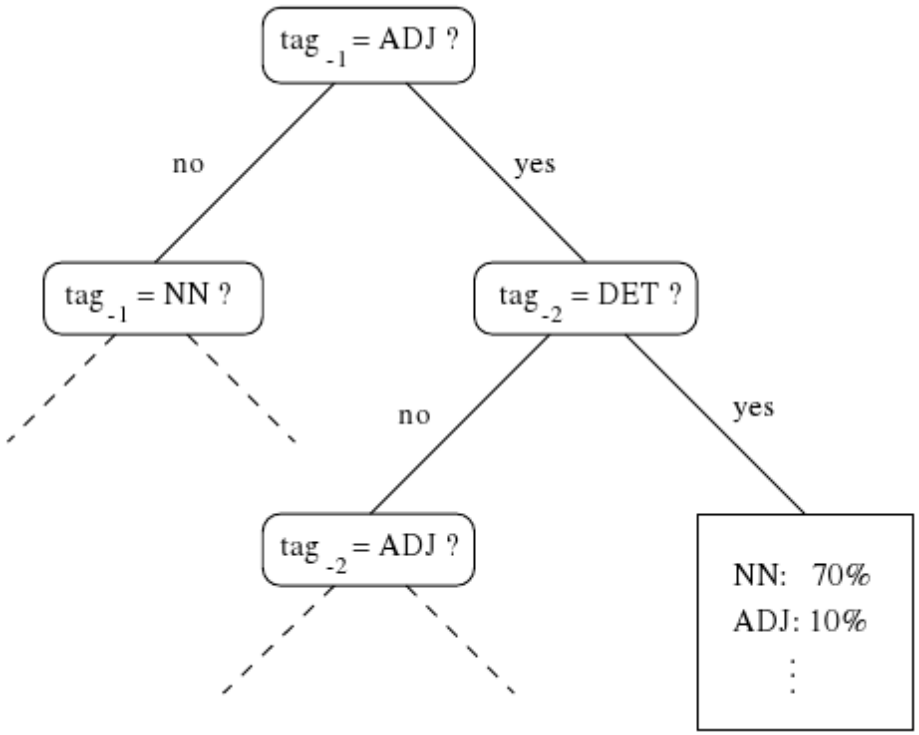
Distinctions

- **Catégorisation (supervisée) :** ranger des objets dans des classes (catégories) prédéfinies
  - catégorisation thématique de documents ;
  - repérage d'entités nommées *(personnes, organisations, lieux)* dans des textes ;
  - détection de relations entre entités *(personne dirige organisation)*.
- **Classification (non supervisée) :** dégager des classes par l'analyse des caractéristiques des objets
  - dégager des classes de clients à partir d'enquêtes de satisfaction clientèle (verbatim des clients : regroupement de textes) ;
  - faire émerger des classes de termes à partir de leurs contextes d'emploi ou de leur constitution.

Quelques formalismes employés en TAL

Arbres de décision

- à chaque embranchement *(nœud)*, prise d'une décision ;
- parcours d'une branche selon le choix ;
- arrêt du parcours aux feuilles terminales.
- exemple : étiquetage morpho-syntaxique



Quelques formalismes employés en TAL

Séparateurs à Vaste Marge (SVM), Noyaux

- Catégorisation de textes
  - Sac de mots : mots communs aux deux textes
  - Noyaux de sous-chaînes : sous-chaînes communes aux deux textes
- Catégorisation de relations
  - Noyaux d'arbres : par exemple, comparaison des arbres élémentaires de deux arbres syntaxiques

Quelques formalismes employés en TAL

Champs conditionnels aléatoires (CRF)

- étiquetage de mots dans une séquence *(suite de mots en tenant compte du contexte phrastique)* ;
  - catégorie morphosyntaxique (Nom, Verbe, Adjectif, ...)
  - syntagme minimal (SN, SP, ...)
  - entité nommée (personne, lieu, organisation, ...)
- repose sur l'étude d'indices :
  - internes inférés du mot *(typographie, taille, désinence, etc.)* ;
  - externes : ressources linguistiques *(dictionnaires, listes)* et outils d'analyse *(sémantique, syntaxique, morphologique)*.

Apprentissage non supervisé

Création de classes

En général ∼ ≪ clustering ≫

Regroupement d'objets

- Induire des classes de mots :
  - catégories morphosyntaxiques pour une langue
    - N, V, A, etc.
  - catégories sémantiques dans un domaine donné
    - médicament, partie du corps, maladie
  - classes thématiques dans des documents
    - économie, sport, médecine
- Induire des classes de documents :
  - classes thématiques (économie, sport, médecine)
  - genres de textes (prose, poésie, théâtre)

Classes distributionnelles

Description des mots par leur contexte

- Détermination du contexte
  - Mots co-occurrents
  - Dépendants syntaxiques
- Méthodes courantes :
  - Vecteurs de contextes
  - Définition d'une distance sur ces vecteurs
  - Regroupement basé sur cette distance
  - Nombreuses variantes selon les choix effectués pour chaque point

Classes distributionnelles

[?]

Induction : optimisation d'un modèle de langue sur les classes

- Supposons les mots d'un corpus répartis dans des classes
- On construit un modèle de langue bigramme sur les classes
  - Séquences de deux classes vues côte à côte
  - On évalue la probabilité d'avoir le corpus observé étant donné ce modèle de langue
- On fait en sorte de répartir les mots dans ces classes de façon à maximiser cette probabilité
  - Revient à maximiser l'information mutuelle moyenne entre deux classes adjacentes
  - L'algorithme proposé produit une classification hiérarchique

Classes thématiques

Modèles à base de thèmes : ≪ Topic models ≫

Modèles génératifs à base de thèmes

- Exploite la matrice mot × document
  - Ensemble de documents formés de mots
  - Ensemble de thèmes
  - Chaque document concerne plusieurs thèmes
- Actuellement : distributions de Dirichlet latentes (LDA)
  - Chaque mot de chaque document est ≪ généré ≫ par
  - choix d'un thème étant donné ce document, puis
  - choix d'un mot étant donné ce thème

Usage des classes induites dans l'apprentissage supervisé

Ajout d'attributs non supervisés dans un catégoriseur supervisé

- Exemple : classes distributionnelles ≪ sémantiques ≫
- Fait typiquement gagner un ou plusieurs points de F-mesure

Mesures d'évaluation

Les mesures d'évaluation

Motivations

- mesurer la progression d'un système en interne ;
- management interne de projet scientifique ;
- comparaison de systèmes *(état de l'art, campagnes d'évaluation)* ;
- objectif marketing *(compréhension d'enquêtes de satisfaction clientèle)*.

Les mesures d'évaluation

- Comparaison :
  - référence : gold standard, réalisé manuellement ;
  - hypothèse : sortie d'un système.
- Matrice de confusion :

		Maladie	
		présente	absente
Test	positif	Vrais Positifs	Faux Positifs
	négatif	Faux Négatifs	Vrais Négatifs

Les mesures d'évaluation

- **référence :** Le sculpteur [personne César] est mort hier à Paris , à l' âge de 77 ans . Comme [personne Yves Montand] , son ami , [personne César Baldaccini] , fils d' émigrés italiens , naît en 1921 dans un quartier très pauvre de Marseille . Le sculpteur , fou de [personne Brancusi] et de [personne Picasso] , se répète beaucoup .
- **hypothèse :** Le sculpteur [personne César] est mort hier à Paris , à l' âge de 77 ans . Comme [personne Yves Montand] , son ami , [personne César Baldaccini] , fils d' émigrés italiens , naît en 1921 dans un quartier très pauvre de [personne Marseille] . Le sculpteur , fou de Brancusi et de Picasso , se répète beaucoup .
- **décomptes :** 3 vrais positifs, 2 faux négatifs, 1 faux positif.
- **problème :** comment définir le nombre de marquables ?

Les mesures d'évaluation

- **rappel (sensibilité) :** taux de vrais positifs.

$$R = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}}$$

- **précision :** valeur prédictive positive.

$$P = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}}$$

- **exactitude :** prédictions justes rapportées au total.

$$Acc = \frac{\text{vrais positifs} + \text{vrais négatifs}}{VP + VN + FP + FN}$$

- spécificité : taux de vrais négatifs.

$$Sp = \frac{\text{vrais négatifs}}{\text{vrais négatifs} + \text{faux positifs}}$$

Les mesures d'évaluation

- **F-mesure :** moyenne harmonique pondérée du rappel et de la précision

$$\text{F-mesure} = \frac{(1+\beta^2)\ x\ \text{Précision}\ x\ \text{Rappel}}{\beta^2\ x\ \text{Précision}\ +\ \text{Rappel}}$$

La valeur attribuée à β permet :

- soit d'équilibrer les poids du rappel et de la précision (β=1) ;
- soit de favoriser :
  - le rappel par rapport à la précison (β=2) ;
  - la précision par rapport au rappel (β=0.5).

Les mesures d'évaluation

- **décomptes :** 3 vrais positifs, 2 faux négatifs, 1 faux positif ;
- **rappel** $= \frac{3}{5} = 0,6$
- **précision** $= \frac{3}{4} = 0,75$
- **F_1-mesure** $= \frac{2 \times 0,6 \times 0,75}{0,6+0,75} = 0,667$

Une question :
Calcul de la significativité statistique

Significativité statistique

- **utilité :** qualifier les différences
  - entre deux systèmes,
  - entre deux configurations d'un même système.
- **difficulté :** comment mesurer la significativité statistique sur les données langagières ?
  - comment formaliser les données ?
    - pour mesurer les différences de catégorisation *(personne, organisation, ville)*,
    - pour mesurer les différences de frontières,
    - pour mesurer ces deux types de différence.
  - peut-on utiliser les valeurs de rappel/précision ?
  - quel test utiliser *(Student, McNemar, etc.)* ?

Significativité statistique

- **difficulté :** comment mesurer la significativité statistique sur les données langagières ?
  - peut-on utiliser un intervalle de confiance ?
    - test de Student (avec $t_\alpha = 1,96$ et risque $\alpha = 0,05$) ?

$$I_c = [\bar{x} - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x} + t_\alpha \frac{\sigma}{\sqrt{n}}] \quad (1)$$

    - simulation de Monte Carlo ?

$$\theta = \lim_{x \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_n := \lim_{x \to \infty} \hat{\theta}_n \quad (2)$$

Dans le TAL, nous ne sommes pas parfaitement outillés pour évaluer tous les phénomènes langagiers.

Nous avons besoin de vos compétences de statisticiens pour utiliser les mesures adéquates qui font sens.


Coates-Stephens, S. (1992).
The analysis and acquisition of proper names for the understanding of free text.
Comput Hum, 26(5) :441–56.


Fleischman, M. (2001).
Automated subcategorization of named entities.
In *Proc of the ACL 2001 Student Research Workshop*, pages 25–30.


Gazdar, G. (1996).
Computing tomorrow.
In Wand, I. and Milner, R., editors, *Paradigm merger in natural language processing*, pages 88–109. Cambridge University Press.


Grishman, R. and Sundheim, B. (1996).
Message understanding conference - 6 : A brief history.

In *Proc of COLING*, pages 466–71, Copenhague, Danemark.

Sekine, S. (2004).
Definition, dictionaries and tagger of extended named entity hierarchy.
In *Proc of LREC*.

Wisniewski, G. (2007).
Apprentissage dans les espaces structurés. Application à l'étiquetage de séquences et à la transformation automatique de documents.
PhD thesis, UPMC.