Text Mining

Slides de Julien Lemoine (Exalead)

Antoine Rozenknop

23 janvier 2009

Présentation

contact

E-Mail: *antoine*(*point*)*rozenknop*(*ad*)*lipn*(*point*)*univ* − *paris*13(*point*)*fr*
Slides: http://www-lipn.univ-paris13.fr/~rozenknop

Plan du cours

Description

1. 1 cours d'introduction au Text Mining (TLN / NLP)
2. 3 cours+tp : Extraction d'information (EI / IE)
3. 3-4 cours+tp : Recherche d'information
4. Applications au LIPN
5. techniques d'apprentissage pour la RI/EI
6. structures de données
7. dernier cours : présentation d'articles

Plan du 1er cours

1. Introduction
2. Niveau lexical
3. Niveau Morphologique/Syntaxique
   - Étiqueteurs grammaticaux
   - Étiqueteurs morpho-syntaxiques
   - Étiqueteurs à base de règles
   - Étiqueteurs probabilistes
   - Stemming
4. Niveau Sémantique et niveau pragmatique
5. Conclusion / Références

Que désigne le Text-Mining

- Des technologies linguistiques permettant de passer du texte (*full text*) à un vecteur numérique (présence-absence ou fréquence). Il est ensuite possible d'appliquer les mêmes algorithmes qu'en Data Mining (par exemple ACP, K-Means, ...)
- Des algorithmes utilisés sur ces vecteurs (par exemple *Naives Bayes Classifier*, *Latent Semantic Analysis/Indexing*, ...)
- Des structures de données adaptées aux propriétés du texte (transducteurs, arbres ternaires, ...)
- Des domaines entiers: *information retrieval*, *information extraction*, ...
- Au final un terme aussi ambigu que Data-Mining !

Que désigne le Text-Mining

- Des technologies linguistiques permettant de passer du texte (*full text*) à un vecteur numérique (présence-absence ou fréquence). Il est ensuite possible d'appliquer les mêmes algorithmes qu'en Data Mining (par exemple ACP, K-Means, ...)
- Des algorithmes utilisés sur ces vecteurs (par exemple *Naives Bayes Classifier*, *Latent Semantic Analysis/Indexing*, ...)
- Des structures de données adaptées aux propriétés du texte (transducteurs, arbres ternaires, ...)
- Des domaines entiers: *information retrieval*, *information extraction*, ...
- Au final un terme aussi ambigu que Data-Mining !

Que désigne le Text-Mining

- Des technologies linguistiques permettant de passer du texte (*full text*) à un vecteur numérique (présence-absence ou fréquence). Il est ensuite possible d'appliquer les mêmes algorithmes qu'en Data Mining (par exemple ACP, K-Means, ...)
- Des algorithmes utilisés sur ces vecteurs (par exemple *Naives Bayes Classifier*, *Latent Semantic Analysis/Indexing*, ...)
- Des structures de données adaptées aux propriétés du texte (transducteurs, arbres ternaires, ...)
- Des domaines entiers: *information retrieval*, *information extraction*, ...
- Au final un terme aussi ambigu que Data-Mining !

Que désigne le Text-Mining

- Des technologies linguistiques permettant de passer du texte (*full text*) à un vecteur numérique (présence-absence ou fréquence). Il est ensuite possible d'appliquer les mêmes algorithmes qu'en Data Mining (par exemple ACP, K-Means, ...)
- Des algorithmes utilisés sur ces vecteurs (par exemple *Naives Bayes Classifier*, *Latent Semantic Analysis/Indexing*, ...)
- Des structures de données adaptées aux propriétés du texte (transducteurs, arbres ternaires, ...)
- Des domaines entiers: *information retrieval*, *information extraction*, ...
- Au final un terme aussi ambigu que Data-Mining !

Que désigne le Text-Mining

- Des technologies linguistiques permettant de passer du texte (*full text*) à un vecteur numérique (présence-absence ou fréquence). Il est ensuite possible d'appliquer les mêmes algorithmes qu'en Data Mining (par exemple ACP, K-Means, ...)
- Des algorithmes utilisés sur ces vecteurs (par exemple *Naives Bayes Classifier*, *Latent Semantic Analysis/Indexing*, ...)
- Des structures de données adaptées aux propriétés du texte (transducteurs, arbres ternaires, ...)
- Des domaines entiers: *information retrieval*, *information extraction*, ...
- Au final un terme aussi ambigu que Data-Mining !

Les besoins en TLN/NLP

- Recherche d'information
- Correction orthographique/grammaticale
- Filtrage/classification d'information : courrier électronique, flux d'actualité, document métier...
- Traduction automatique
- Résumé automatique
- Question/réponse (interfaces en langage naturel)
- ...

Les besoins en TLN/NLP

- Recherche d'information
- Correction orthographique/grammaticale
- Filtrage/classification d'information : courrier électronique, flux d'actualité, document métier...
- Traduction automatique
- Résumé automatique
- Question/réponse (interfaces en langage naturel)
- ...

Les besoins en TLN/NLP

- Recherche d'information
- Correction orthographique/grammaticale
- Filtrage/classification d'information : courrier électronique, flux d'actualité, document métier...
- Traduction automatique
- Résumé automatique
- Question/réponse (interfaces en langage naturel)
- ...

Les besoins en TLN/NLP

- Recherche d'information
- Correction orthographique/grammaticale
- Filtrage/classification d'information : courrier électronique, flux d'actualité, document métier...
- Traduction automatique
- Résumé automatique
- Question/réponse (interfaces en langage naturel)
- ...

Les besoins en TLN/NLP

- Recherche d'information
- Correction orthographique/grammaticale
- Filtrage/classification d'information : courrier électronique, flux d'actualité, document métier...
- Traduction automatique
- Résumé automatique
- Question/réponse (interfaces en langage naturel)
- ...

Les besoins en TLN/NLP

- Recherche d'information
- Correction orthographique/grammaticale
- Filtrage/classification d'information : courrier électronique, flux d'actualité, document métier...
- Traduction automatique
- Résumé automatique
- Question/réponse (interfaces en langage naturel)
- ...

Positionnement IR/IE

- **Recherche d'informations (IR)**: identifie un ensemble de documents à partir d'un ensemble plus large (document assimilé à un "sac de mots").
  Ex: *Trouver les documents qui traitent de rachats d'entreprises*
- **Extraction d'informations (IE)**: extrait et structure de l'information précise contenue dans un document.
  Ex: *Etablir une base de données où l'on peut retrouver les noms des entreprises informatiques cédées en 2003*
- **Compréhension de texte**: représente de façon explicite toute l'information d'un document (rhétorique, intentionnalité, ...)
  Ex: *Déterminer les différentes visées stratégiques sous-jacents à travers ces ventes et acquisitions*

Opposition Naturel/Formel

- les langages formels sont par construction **explicites** et **non ambigus**
- les langages naturels sont par essence **implicites** et **ambigus**
- **implicite**:
  - Je préfère Emacs à Vim car je **le** connaîs bien et **il** me permet de gagner du temps
  - Je viens de trouver un bug, mon voisin **aussi**
- **ambigu**:
  - la belle ferme le voile
  - les experts ont $\binom{analysé}{dissocié}$ **la croissance de la consommation**

Opposition Naturel/Formel

- les langages formels sont par construction **explicites** et **non ambigus**
- les langages naturels sont par essence **implicites** et **ambigus**
- **implicite**:
  - Je préfère Emacs à Vim car je **le** connaîs bien et **il** me permet de gagner du temps
  - Je viens de trouver un bug, mon voisin **aussi**
- **ambigu**:
  - la belle ferme le voile
  - les experts ont $\binom{analysé}{dissocié}$ **la croissance de la consommation**

Opposition Naturel/Formel

- les langages formels sont par construction **explicites** et **non ambigus**
- les langages naturels sont par essence **implicites** et **ambigus**
- **implicite**:
  - Je préfère Emacs à Vim car je **le** connaîs bien et **il** me permet de gagner du temps
  - Je viens de trouver un bug, mon voisin **aussi**
- **ambigu**:
  - la belle ferme le voile
  - les experts ont $\binom{analysé}{dissocié}$ **la croissance de la consommation**

Besoin de ressources linguistiques

- La plupart des systèmes utilisent des ressources linguistiques
- les compétences linguistiques sont rarement présentes dans l'industrie
- les ressources linguistiques sont souvent difficiles (et donc chères) à produire
- les ressources sont souvent au moins aussi coûteuses que la technologie elle-même

Historique

- jusqu'à la fin des années 80: principalement des modèles essayant de formaliser notre compréhension du langage de manière formelle (approche généraliste)
- depuis les années 90: application à un domaine donné en utilisant des modèles extraits automatiquement/semi-automatiquement depuis des corpus textuels. Émergence de technologies généralistes de traitement de la langue

Les différentes couches

- niveau morpho-lexical: reconnaître les “mots” (tokenization)
- niveau syntaxique: étiqueter les séquences de mots
- niveau sémantique: comprendre (arêtes / manger du poisson)
- niveau pragmatique : contextualiser (il a demandé l'addition à la serveuse et l'a payée)

Exemple de texte à transformer en descripteurs

Un accord Microsoft-Yahoo peut-être ce week-end
[2008-05-03 09:35]
NEW YORK (Reuters) - Microsoft et Yahoo négocient et leurs discussions se sont même intensifiées pour tenter de parvenir à un accord amiable par lequel le groupe internet passerait dans le giron du numéro un mondial des logiciels, a-t-on indiqué vendredi de sources proches du dossier.

Plan

1. Introduction
2. **Niveau lexical**
3. Niveau Morphologique/Syntaxique
   - Étiqueteurs grammaticaux
   - Étiqueteurs morpho-syntaxiques
   - Étiqueteurs à base de règles
   - Étiqueteurs probabilistes
   - Stemming
4. Niveau Sémantique et niveau pragmatique
5. Conclusion / Références

Définition

- reconnaître les mots: qu'est-ce qu'un mot ?
- Langues non alphabétiques (chinois), Langues sans séparateurs (thaïlandais)
- Quels séparateurs ?
  - mots composés: parce que, pomme de terre
  - apostrophe: Aujourd'hui, qu'il, 1'000 euros
  - point: E.U., 3.2 degrés, lib.net, fichier .so
  - trait d'union: grand-père, est-ce, Croix-Rouge, Jean-Louis
  - underscore: login_n, _dl_out_of_memory
  - et commercial: C&A, R&D, D&CO
- Domaine privilégié de la **morphologie** (étude de la structure des mots) et de la **lexicographie** (recensement et classification des formes d'une langue)
- **Représentation: lexique important pour chaque langue**

Niveau lexical - Exemple

Un accord Microsoft-Yahoo peut-être ce week-end [2008-05-03 09:35]
NEW YORK (Reuters) - Microsoft et Yahoo négocient et leurs discussions se sont même intensifiées pour tenter de parvenir à un accord amiable par lequel le groupe internet passerait dans le giron du numéro un mondial des logiciels, a-t-on indiqué vendredi de sources proches du dossier.

Niveau syntaxique

Définition

- Objectif: affecter une catégorie grammaticale à chaque mot (NC, NP, V, Det, ...)
- Ambiguité : nous **avions** déjà vu ces **avions**
- Deux niveaux de ressources : des lexiques donnant les mots ainsi que leurs catégories grammaticales possibles
- Un système de désambiguïsation par apprentissage (en utilisant un corpus annoté manuellement)

Exemple de Lexique avec étiquettes

```
avions.N
avions.V
ferme.N
ferme.V
ferme.A
belle.A
belle.N
```

Plan

1. Introduction
2. Niveau lexical
3. **Niveau Morphologique/Syntaxique**
   - Étiqueteurs grammaticaux
   - Étiqueteurs morpho-syntaxiques
   - Étiqueteurs à base de règles
   - Étiqueteurs probabilistes
   - Stemming
4. Niveau Sémantique et niveau pragmatique
5. Conclusion / Références

Étiqueteurs grammaticaux

Définition

- but: associer une étiquette grammaticale (ou tag) à chaque mot d'une phrase (ADJ, NC, NP, V, Det).
- On parle également de Part Of Speech (PoS) Tagging/Tagger

Pourquoi?

- Plus facile que de comprendre une phrase (taux de bon étiquetage $> 95\%$)
- Mais utile dans de nombreuses applications :
  - extraction d'information
  - moteur de recherche
  - classification supervisée/non supervisée
  - ...
- Les POS suffisent souvent à identifier des groupes syntaxiques simples comme les groupes nominaux

Exemple

la séance est ouverte à 15h43.

phrase	étiquettes
la	Det+Feminin+Singulier
séance	NomCommun+Feminin+Singulier
est	Verbe+IndPré+Singulier+3Pers
ouverte	Verbe+ParPas+feminin+Singulier
à	Prep
15	Num
h	NC
43	Num
.	Ponc

Est-il difficile d'étiqueter ?

la belle ferme le voile

ART NC V ART NC : s'il s'agit d'une femme qui ferme un voile

la belle ferme le voile

ART ADJ NC PRO V : si une ferme voile la vue d'autre chose

Ambiguité

```
ferme,fermer.V:P1s:P3s (verbe)
ferme,.N:fs (nom commun féminin singulier)
ferme,.A:ms:fs (adjectif singulier)
belle,beau.A:fs (adjectif féminin singulier)
belle,.N:fs (nom commun féminin singulier)
```

Est-t-il difficile d'étiqueter ?

Déterministe

- Il existe également de nombreux mots qui ne sont étiquetables que par une seule étiquette :

  ```
  âge,.N:ms (nom commun masculin singulier)
  éducatif,.A:ms (adjectif masculin singulier)
  électoraux,électoral.A:mp (adjectif masculin singulier)
  aurez,avoir.V:F2p (verbe)
  ```

- peut dépasser 50% des mots d'un grand corpus (dépend beaucoup du corpus et de la langue)

Quelle information utiliser ?

Certaines séquences sont plus fréquentes

- il est plus fréquent d'avoir la séquence **ART ADJ NC** (ex: le gentil homme) que **ART ADJ V** (ex: la favorite équipe son cheval...)
- Un étiqueteur qui se baserait sur cette information devrait associer l'étiquette NC à *ébauche* plutôt qu'une étiquette V dans la phrase :

  `une belle ébauche la montre`

- il faut pour cela s'appuyer sur le contexte du mot que l'on souhaite étiqueter
- le contexte peut être faux (belle pourrait avoir l'étiquette NC)
- En pratique, cette information ne suffit pas (taux de 77%)

Quelle information utiliser ?

La nature même du mot

- *belles* est probablement plus fréquemment employé en français comme un adjectif que comme un nom commun
- En fait, Charniak (1993) a montré qu'un étiqueteur simple qui étiquette un mot par son étiquette la plus fréquente permet d'obtenir des taux d'étiquetage de l'ordre de 90%
- pour obtenir cette information: il faut un corpus étiqueté

Taux d'étiquetage

Est-ce qu'un taux de 95% est un bon taux ?

- 5 erreurs tous les 100 mots
- si on considère une phrase = 20 mots
- ceci fait donc potentiellement une erreur par phrase (en pratique, plusieurs erreurs peuvent intervenir dans la même phrase)
- tout dépend de l'application, il est difficile de comparer des étiqueteurs entraînés sur des corpus différents (% de mots ayant plusieurs étiquettes, taille du vocabulaire, jeu d'étiquettes, taux des inconnus à l'apprentissage, ...)

Niveau morpho-syntaxique: Outils

Buts

- résolution de certaines ambiguïtés (e.g. avions.V/avions.N)
- réduction de la taille du vocabulaire (uniquement des lemmes)
- suppression de la variabilité lexicale

Outils

- Étiquetage morpho-syntaxique
- Lemmatisation
- Stemming

Plan

1. Introduction
2. Niveau lexical
3. **Niveau Morphologique/Syntaxique**
   - Étiqueteurs grammaticaux
   - Étiqueteurs morpho-syntaxiques
   - Étiqueteurs à base de règles
   - Étiqueteurs probabilistes
   - Stemming
4. Niveau Sémantique et niveau pragmatique
5. Conclusion / Références

Étiquetage morpho-syntaxiques

Définition

Affectation automatique, en fonction du contexte, d'étiquettes morpho-syntaxiques

Exemple

Au	Kremlin-Bicêtre	la	petite	brise	est	fraîche
Prep	NP	Det	Adj	NC	V	Adj

Non trivial car

- **ambiguïtés**: petite:NC, brise:V, fraîche:NC
- **mot inconnu**: Kremlin-Bicêtre

Étiquetage morpho-syntaxique

Formalisation

Étant donné un ensemble de couples (mot, étiquette) et un texte, choisir pour chacun des mots du texte parmi ces étiquettes associées celles qui correspondent au **contexte**

Plusieurs approches

- A base de règles: l'*étiqueteur de Brill*
- Probabiliste: Chaînes de Markov cachées (HMM)
- Probabiliste: arbres décisionnels (TreeTagger), SVM, ...

Lemmatisation

Définition

- Réduction automatique des mots à une forme canonique
- Forme canonique= infinif pour les verbes, singulier+masculin pour les noms, ...
- Exemple:
  - chats $\rightarrow$ chat
  - mangerons $\rightarrow$ manger
- facile si on a un mécanisme de désambiguïsation (avions.V vs avions.NC)

Plan

1. Introduction
2. Niveau lexical
3. **Niveau Morphologique/Syntaxique**
   - Étiqueteurs grammaticaux
   - Étiqueteurs morpho-syntaxiques
   - Étiqueteurs à base de règles
   - Étiqueteurs probabilistes
   - Stemming
4. Niveau Sémantique et niveau pragmatique
5. Conclusion / Références

Étiquetage à base de règles

Étiqueteur de Brill

- Eric Brill : thèse 1993
- principe: "error-driven transformation-based"
- error-driven: apprentissage supervisé
- transformation-based:
  - L'étiquette la plus fréquente du mot est affectée
  - application d'un ensemble de règles pour changer les étiquettes jusqu'à stabilisation

Phases

1. apprentissage: une fois pour toutes, lent, complexe
2. utilisation: souvent, rapide, simple

Étiqueteur de Brill

Algorithme: initialisation

- pour les mots connus (dans le lexique): étiquette la plus probable
- pour les mots inconnus
  - (1992) NP pour les mots avec majuscule, NC pour les autres
  - (1994) Apprentissage de “guessing rules”

Algorithme: Application

On applique toutes les règles de la base

Étiqueteur de Brill

Algorithme: apprentissage

- Calcule de façon itérative le "score" de chaque règle proposée, i.e. le nombre d'erreurs avant application de la règle moins le nombre d'erreurs après son application
- Choisit la meilleure règle (score le plus haut), l'ajoute à la base de règle et l'applique effectivement au texte
- Répète jusqu'à ce qu'aucune règle proposée n'ait un score suffisant (seuil)
- D'où viennent les règles ?

Étiqueteur de Brill

deux types de règles

- Pour affecter une étiquette à un mot inconnu lors de l'initialisation (lexical rules)
- Pour corriger (dans un contexte) une étiquette existante (contextual rules)

Forme des règles

- Lexicales : mot $\longrightarrow$ étiquette si Condition
- Contextuelles: $\textit{étiquette}_1 \longrightarrow \textit{étiquette}_2$ si Condition

Création des règles

- On part d'un corpus étiqueté à la main (minimum une centaine de pages)
- Création d'un dictionnaire comportant tous les mots ainsi que leurs étiquettes (l'étiquette la plus fréquente en tête de liste)
- Le corpus est chargé et on affecte l'étiquette la plus fréquente dans le dictionnaire à chaque mot
- A chaque erreur rencontrée, des règles sont créées (modèles)

Exemple de conditions

Règles lexicales

- le mot courant a le suffixe x
- le mot courant a le préfixe x
- supprimer le préfixe/suffixe au mot courant conduit à un mot inconnu
- le mot courant apparaît après/avant le mot m' (fixé pour une règle donnée)
- le mot courant contient le caractère x

Règles contextuelles

- (une, deux, trois) étiquette(s) précédente(s)/suivante(s) est(sont) X
- le bigramme d'étiquettes précédentes/suivantes est XY
- le(s) mot(s) précédents/suivants est(sont) W
- le mot courant est W et le mot suivant/précédent est W'
- le mot courant est W et l'étiquette précédente/suivante est X

Exemple de règles

Règles lexicales

- **SBC:sg e fdeletesuf 1 ADJ:sg**: Si le mot est catégorisé substantif singulier et qu'en enlevant le suffixe e, on obtient un mot existant, changer son étiquette pour adjectif singulier.
- **SBC:pl nous fgoodright VCJ:pl**: Si le mot est catégorisé substantif pluriel et qu'il est à droite de *nous*, changer son étiquette pour verbe conjugué pluriel.
- **SBC:pl ais fhasssuf 3 VCJ:sg**: Si le mot est catégorisé substantif pluriel et qu'il comporte le suffixe *ais*, changer son étiquette pour verbe conjugué singulier

Exemple de règles

Règles contextuelles

- **DTN:sg PRV:sg NEXTTAG VCJ:sg**: Si le mot est catégorisé déterminant singulier, qu'il présente une ambiguïté déterminant sg./pronom verbal sg. et que le mot suivant est catégorisé verbe conjugué sg., changer son étiquette pour pronom verbal sg. (ex: je **le** juge)
- **SBC:sg ADJ:sg PREVTAG SBC:sg**: Si le mot est catégorisé substantif sg., qu'il présente une ambiguïté sg./adjectif et que le mot précédent est catégorisé substantif sg., changer son étiquette pour adjectif sg. (ex: la maison **or**)
- **PRV:sg PRO:sg WDPREVTAG PREP elle**: si le mot est catégorisé pronom verbal sg., qu'il présente une ambiguïté pronom verbal sg./pronom sg. et que le mot précédent est *elle* ou une préposition, changer son étiquette pour pronom sg. (ex: on parle de **lui**)

Plan

1. Introduction
2. Niveau lexical
3. **Niveau Morphologique/Syntaxique**
   - Étiqueteurs grammaticaux
   - Étiqueteurs morpho-syntaxiques
   - Étiqueteurs à base de règles
   - Étiqueteurs probabilistes
   - Stemming
4. Niveau Sémantique et niveau pragmatique
5. Conclusion / Références

Principe

- Soit $W_1^n = w_1 \dots w_n$ une séquence de n mots
- Étiqueter W_1^n revient à chercher la séquence de catégories grammaticales $C_1^n = C_1 \dots C_n$ telle que la probabilité conditionnelle $P(C_1, \dots, C_n \mid w_1, \dots, w_n)$ soit maximale.
- Comment trouver

$$\tilde{C_1^n} = \underset{C_1^n}{\operatorname{argmax}} P(C_1^n \mid W_1^n)$$

- Loi de Bayes:

$$P(C_1^n \mid W_1^n) = \frac{P(W_1^n \mid C_1^n).P(C_1^n)}{P(W_1^n)}$$

Principe

- La maximisation étant faite à séquence W_1^n, **fixée** :

$$\underset{C_1^n}{\operatorname{argmax}}\, P(C_1^n \mid W_1^n) = \underset{C_1^n}{\operatorname{argmax}}(P(W_1^n \mid C_1^n) \times P(C_1^n))$$

- De plus :

$$\begin{aligned} P(W_1^n \mid C_1^n) &= P(w_1 \mid C_1^n) \times P(w_2 \mid w_1, C_1^n) \times \\ &\quad P(w_3 \mid W_1^2, C_1^n) \times \\ &\quad \ldots \times P(w_n \mid W_1^{n-1}, C_1^n) \end{aligned}$$

$$\begin{aligned} P(C_1^n) &= P(C_1) \times P(C_2 \mid C_1) \times P(C_3 \mid C_1^2) \times \\ &\quad \ldots \times P(C_n \mid C_1^{n-1}) \end{aligned}$$

Étiquetage morpho-syntaxique par HMM

Hypothèses

- Conditionnement lexical limité :

$$P(w_i \mid \ldots, C_i, \ldots) = P(w_i \mid C_i)$$

- Horizon de dépendance syntaxique limité: k voisins

$$P(C_i \mid C_1, \ldots, C_{i-1}) = P(C_i \mid C_{i-k}, \ldots, C_{i-1})$$

Étiquetage morpho-syntaxique par HMM

Initialement

$$
\begin{aligned}
P(W_1^n \mid C_1^n) &= P(w_1 \mid C_1^n) \times P(w_2 \mid w1, C_1^n) \times \\
&\quad \ldots \times P(w_n \mid W_1^{n-1}, C_1^n) \\
P(C_1^n) &= P(C_1) \times P(C_2|C_1) \times \ldots \times P(C_n|C_1^{n-1})
\end{aligned}
$$

D'où

$$
\begin{aligned}
P(W_1^n \mid C_1^n) &= P(w_1 \mid C_1) \times \ldots \times P(w_n \mid C_n) \\
P(C_1^n) &= P(C_1^k) \times P(C_{k+1} \mid C_1, \ldots, C_k) \times \ldots \\
&\quad \times P(C_n \mid C_{n-k}, \ldots, C_{n-1})
\end{aligned}
$$

Finalement

$$
\begin{aligned}
P(W_1^n \mid C_1^n) \times P(C_1^n) &= P(W_1^k \mid C_1^k) \times P(C_1^k) \times \\
&\quad P(W_{k+1}^n \mid C_{k+1}^n) \times P(C_{k+1}^n) \\
&= P(W_1^k \mid C_1^k) \times P(C_1^k) \times \\
&\quad \prod_{i=k+1}^{i=n} (P(w_i \mid C_i) \times P(C_i \mid C_{i-k}^{i-1}))
\end{aligned}
$$

Le modèle peut être interprété comme une chaîne de Markov cachée d'ordre k

Définition

- processus stochastique ξ sur l'espace d'états $C = \{C_1, \ldots, C_n\}$ possédant la propriété markovienne:

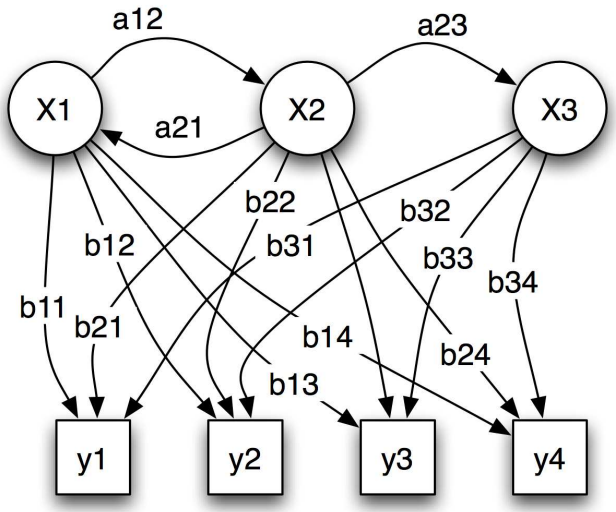
$$P(\xi_t \mid \xi_1, \ldots, \xi_{t-1}) = P(\xi_t \mid \xi_{t-k}, \ldots, \xi_{t-1})$$

- $k =$ ordre de la chaîne de Markov
- En pratique, $k = 1$ (bigrammes) ou 2 (trigrammes), > 2 rare
- Vocabulaire:

$$P(\xi_1, \ldots, \xi_t) = P(\xi_1) \times P(\xi_2 \mid \xi_1) \times P(\xi_t \mid \xi_{t-1})$$

- avec $P(\xi_1) =$ probabilités initiales
- et $P(\xi_2 \mid \xi_1), \ldots, P(\xi_t \mid \xi_{t-1}) =$ probabilités de transitions

Modèle de Markov caché



Définition

- Un ensemble d'états $X = \{X_1, \ldots, X_n\}$
- Un ensemble de transitions : matrice A tel que $a_{ij} = P(X_j \mid X_i)$
- Un ensemble d'observations $Y = \{y_1, \ldots, y_m\}$
- Les probabilités d'observations : matrice B tel que $b_{ij} = P(y_j \mid X_i)$
- un vecteur initial de probabilité I tel que $I_i = P_I(X_i)$

Exemple de HMM

Exemple simple

- une personne éloignée dont on connaît les activités (Balade, Code, TV), on veut en déduire le temps chez lui (Soleil, Pluie)
- On connaît les probabilités de Pluie/Soleil temps globales ($I_{pluie} = 60\%, I_{soleil} = 40\%$)
- On connaît la matrice de probabilités A :

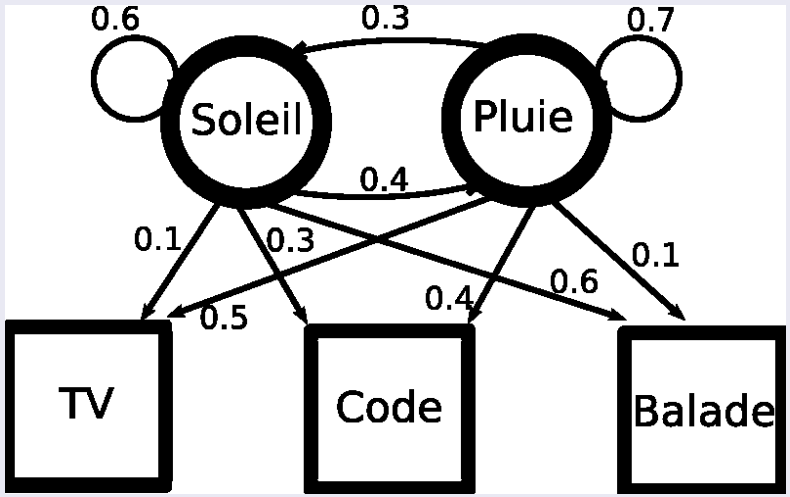
	Pluie	Soleil
Pluie	0.7	0.3
Soleil	0.4	0.6

- On connait la matrice de probabilités B :

	balade	Code	TV
Pluie	0.1	0.4	0.5
Soleil	0.6	0.3	0.1

Exemple de HMM

Exemple simple



Quel est le temps en sachant ses activités ?

	J1	J2	J3	J4
Activité	Balade	Code	Balade	TV

Exemple de HMM

Exemple simple

Quel est le temps en sachant ses activités ?

	J1	J2	J3	J4
Activité	Balade	Code	Balade	TV

Solution

Pour étiqueter une suite d'activités $a_1, \ldots, a_n$, il faut trouver le chemin dans l'automate permettant l'observation de $a_1, \ldots, a_n$ et maximisant la probabilité totale:

$$\prod_{i=1}^{n} (P(a_i \mid temps_i) \times P(temps_i \mid temps_{i-1})$$

HMM comme étiqueteur

- Les états X sont les étiquettes (V, N, ADJ, ...)
- Les observations sont les mots

$$a_{ij} = P(PoS_j \mid PoS_i) = \frac{nb\ occ.\ PoS_i\ suivi\ PoS_j}{nb\ occ.\ PoS_i}$$

$$b_{ij} = P(w_j \mid PoS_i) = \frac{nb\ occ.\ w_j\ étiquetées\ PoS_i}{nb\ mots\ étiquetés\ PoS_i}$$

$$I_i = P_I(X_i) = \frac{nb\ mots\ étiquetés\ PoS_i}{nb\ mots}$$

Application

- Pour étiqueter une suite de mots $w_1, \ldots, w_n$, il faut trouver le chemin dans l'automate permettant l'observation de $w_1, \ldots, w_n$ et maximisant la probabilité totale:

$$\prod_{i=1}^{n}(P(w_i \mid PoS_i) \times P(PoS_i \mid PoS_{i-1})$$

- Algorithme rapide: Viterbi

HMM: Extensions

Mots inconnus

- Dans un grand corpus, il y a beaucoup de mots inconnus
- Performances dépendent énormément des étiquettes données aux mots inconnus
- Solution: utiliser des informations morphologiques (Majuscules, suffixes, césure, ...). Même principe que l'étiqueteur de Brill (guesser)

Trigrammes

- Plutôt que d'utiliser l'étiquette actuelle et l'étiquette précédente (bigramme), on utilise l'étiquette actuelle et les deux précédentes
- Donne un meilleur modèle des dépendances linguistiques entre étiquettes
- Problème: beaucoup de trigrammes n'apparaissent pas dans le corpus d'apprentissage: utilisation d'interpolation linéaire (ou autre) entre deux bigrammes pour construire un trigramme

Algorithme de Viterbi (1/2)

Comment trouver la meilleure séquence d'étiquettes rapidement ?

- utilisé pour trouver efficacement $O(W \times n^2)$ la séquence d'états cachés (chemin de Viterbi) maximisant la probabilité des événements observés. W = nombres de mots
- Principe: programmation dynamique. Deux tableaux T et U, T contient la meilleure solution à l'instant t et U contient la meilleure solution à l'instant $t+1$
- T et U contiennent n éléments ayant chacun deux types d'informations:
  - vpath: le chemin de Viterbi constitué de t éléments pour atteindre l'état X_i
  - vprob: la probabilité associée au chemin de Viterbi (la meilleure probabilité pour atteindre l'état X_i à l'instant t)

Algorithme de Viterbi (2/2)

Init: $T_i = \{vpath = \{X_i\}, vprob = I_i\}$
for $(w_i \in W)$
  $U = []$
  for $(nextState \in X)$
    $bprob = 0$
    $bpath = []$
    for $(sourceState \in X)$
      $p = b[sourceState][w_i] \times a[sourceState][nextState]$
      $tmp = T[source_state]$
      $tmp.vprob* = p$
      if $(tmp.vprob > bprob)$
        $bprob = tmp.vprob$
        $bpath = tmp.vpath + [nextState]$
    $U[nextState] = \{vpath = bpath, vprob = bprob\}$
retourne le chemin ayant le plus grand $U[state].vprob$

Estimation des paramètres du modèle

supervisé

- corpus de phrases étiquetées à la main
- calcul direct des valeurs
- Problème des données manquantes

non supervisé

- corpus de phrases non étiquetées
- Algorithme Baum-Welch (mise à jour des probabilités)
- Forte **sensibilité aux paramètres initiaux**

Méthodes hybrides

- On initialise par des paramètres estimés sur un corpus annoté
- On applique un algorithme Baum-Welch sur un important ensemble de phrases non étiquetées

Conclusion

- Algorithmes efficaces (Viterbi)
- Possibilités d'apprentissage
- Performances: 95% à 98%

Ressources disponibles

- XRCE: étiqueteur morpho-syntaxique par HMM en ligne, trois sorties possibles : Tokenization, Analyse morphologique et enfin désambiguïsation via HMM: http://www.xrce.xerox.com/competencies/content-analysis/demos/french
- Exemple de corpus téléchargable (dépêches en allemand): http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html
- Liste des étiqueteurs disponibles (pas très à jour): http://www-nlp.stanford.edu/links/statnlp.html

Étiquetage morpho-syntaxique - Exemple

Un accord Microsoft-Yahoo peut-être ce week-end
[2008-05-03 09:35]
NEW YORK (Reuters) - Microsoft et Yahoo négocient et leurs discussions se sont même intensifiées pour tenter de parvenir à un accord amiable par lequel le groupe internet passerait dans le giron du numéro un mondial des logiciels, a-t-on indiqué vendredi de sources proches du dossier.

Nom propre	Adverbe
Nom commun	Verbe
Adjectif	Pronom
	Nombre
	Préposition
	Ponctuation
	Déterminant
	Conjonction

Étiquetage morpho-syntaxique - Exemple

Liste des descripteurs

lemme	étiquette	lemme	étiquette
accord	nom commun	Microsoft-Yahoo	nom propre
week-end	nom commun	NEW	nom propre
YORK	nom propre	Reuters	nom propre
Microsoft	nom propre	Yahoo	nom propre
discussion	nom commun	accord	nom commun
amical	adjectif	groupe	nom commun
internet	adjectif	giron	nom commun
numéro un	nom commun	mondial	adjectif
logiciel	nom commun	vendredi	nom commun
source	nom commun	proche	adjectif
dossier	nom commun		

Étiquetage morpho-syntaxique - Exemple

Liste des descripteurs (avec tri-grams)

lemme	étiquette	lemme	étiquette
accord	NC	accord amiable	GN
Microsoft-Yahoo	NP	groupe internet	GN
week-end	NC	giron numéro un mondial	GN
NEW YORK	NP	numéro un mondial logiciel	GN
Reuters	NP	vendredi source proche	GN
Microsoft	NP	source proche dossier	GN
Yahoo	NP		
discussion	NC		

Plan

1. Introduction
2. Niveau lexical
3. **Niveau Morphologique/Syntaxique**
   - Étiqueteurs grammaticaux
   - Étiqueteurs morpho-syntaxiques
   - Étiqueteurs à base de règles
   - Étiqueteurs probabilistes
   - Stemming
4. Niveau Sémantique et niveau pragmatique
5. Conclusion / Références

Stemming

Définition

- Principe: Supprimer le suffixe d'un mot jusqu'à obtenir une racine.
- Cette racine est appelée **stem**
- Algorithmes très souvent basés sur des règles
- Par exemple les mots continu, continua, continuait, continuant sont tous ramenés au stem **continu**, "fish", et "fisher" doivent être ramenés à la racine "fish".
- Algorithme le plus connu: **Porter stemmer**
- Langage de règles libres: snowball avec de nombreuses bases de règles : anglais, francais, espagnol, portugais, italien, allemand, ...
- http://snowball.tartarus.org

Extrait des règles fr de snowball

```
define i_verb_suffix as setlimit tomark pV for (
  [substring] among (
    'iraient' 'irais' 'irait' 'iras' 'irent' 'irez' 'iriez'
    'irions' 'irons' 'iront' 'is' 'issaient' 'issais' 'issait
    'issant' 'issante' 'issantes' 'issants' 'isse' 'issent' 'i
    'issez' 'issiez' 'issions' 'issons' 'it'
      (non-v delete)
  )
)
```

Recherche les suffixes de verbes commençants par **i**, supprime le suffixe si il est précédé d'une consonne *non-v = non voyelle*

Stemming

Exemple

un accord microsoft-yahoo peut-êtr ce week-end
2008-05-03 09:35
new york reuter microsoft et yahoo négocient et leur discuss se sont mêm intensifi pour tent de parven à un accord amiabl par lequel le group internet pass dan le giron du numéro un mondial de logiciel indiqu vendred de sourc proch du dossi

Plan

1. Introduction
2. Niveau lexical
3. Niveau Morphologique/Syntaxique
   - Étiqueteurs grammaticaux
   - Étiqueteurs morpho-syntaxiques
   - Étiqueteurs à base de règles
   - Étiqueteurs probabilistes
   - Stemming
4. **Niveau Sémantique et niveau pragmatique**
5. Conclusion / Références

Les niveaux sémantique et pragmatique

Définition

- Sémantique: étude du sens hors contexte (i.e. le sens littéral). Exemple détection d'entités nommées ou de groupes nominaux
- Pragmatique: étude du sens en contexte (co-références), très utile en extraction d'informations

Définitions

- Représentation de la connaissance sous forme d'arbre Hyperonymie(est-un)/Hyponymie(spécialisation) ou de Graphe
- Les ambiguïtés sémantiques (moins fréquentes que les ambiguïtés grammaticales) sont résolues par le contexte
- Bien adapté pour représenter les connaissances d'un domaine (d'une société par exemple) mais difficile à étendre pour une approche généraliste
- Utilisé pour la catégorisation, classification non supervisée, la recherche d'information (veille économique par exemple), ...

Détection

Deux grands types d'approches de détection des entités sémantiques dans le texte:

- Basées sur des dictionnaires: un ensemble d'expressions (full-text, lemmatisé, ...) sont stockées dans un dictionnaire avec des étiquettes sémantiques
- Basées sur des grammaires (transducteurs): un linguiste modélise l'entité sémantique à extraire (par exemple un Lieu)

Ambiguïté semantique

- il y a toujours des ambiguïtés (moins qu'au niveau morphologique/syntaxique)
- Exemple d'ambiguïté:
  Orange en tant que ville, fruit, couleur, société, ...
- La plupart du temps, résolu en utilisant le contexte:
  le plus souvent une liste de mots (ngrams) qui permet de déterminer le bon sens lorsqu'ils sont dans la même phrase
- Exemple de résolution d'ambiguïté par le contexte: avocat
  si le contexte contient manger $\longrightarrow$ avocat(fruit)
  si le contexte contient juge $\longrightarrow$ avocat(métier)

Exemple de dictionnaire sémantique

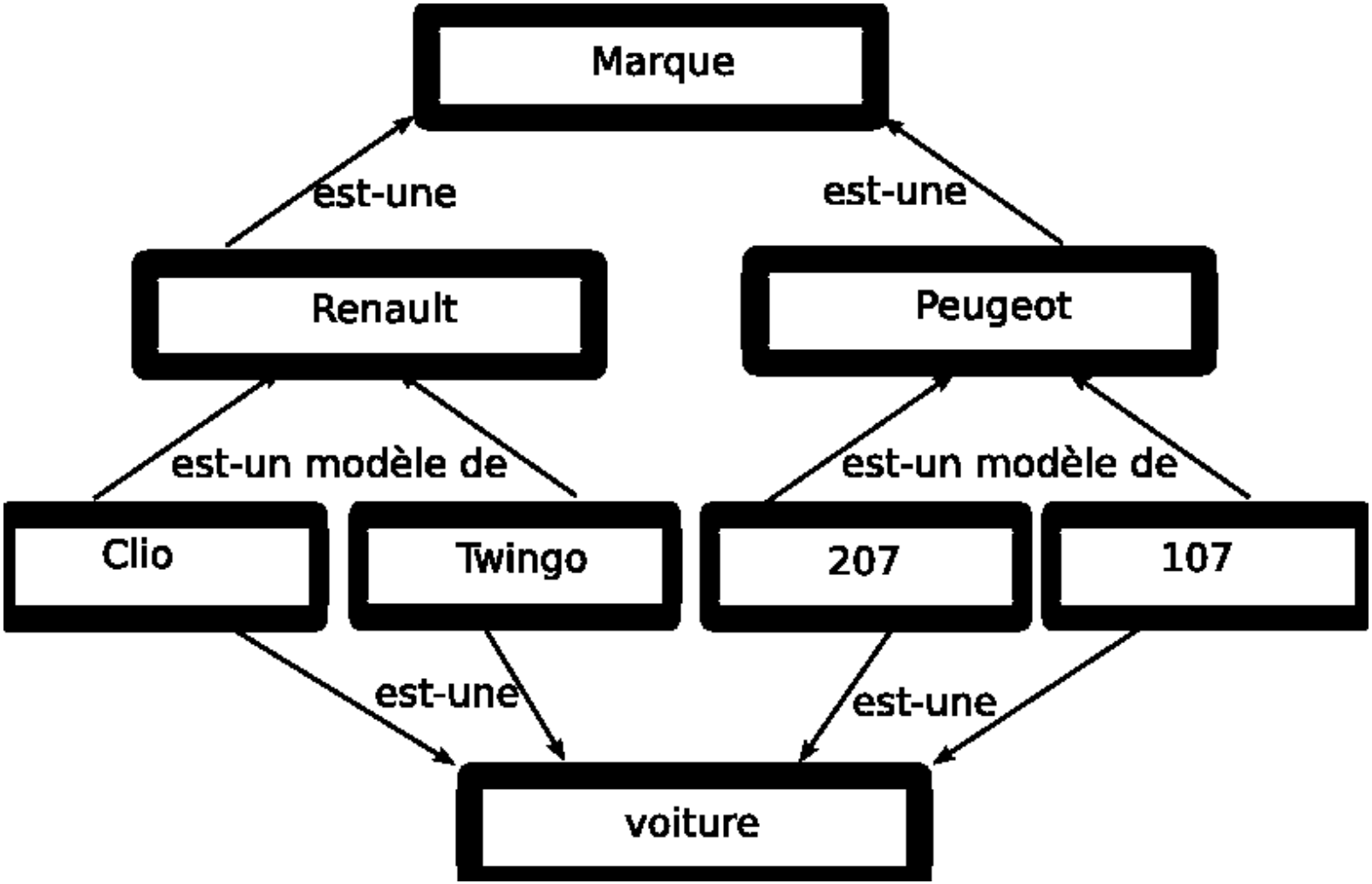


Figure: Exemple de dictionnaire sémantique du domaine automobile

Utilisation des informations sémantiques

Cas d'utilisation

- enrichissement des moteurs de recherches
- extension de requêtes
- facilité de lecture de documents (sur-lignage des informations sémantiques)
- veille (sentiments, entités nommées, produits, marques, ...)
- ...

Principe

- Organiser l'information lexicale en terme de signification de mots
- Approximativement 95600 formes de mots et 70100 groupes de synonymes (synsets) pour l'anglais
- Formes des mots : désignation de l'expression physique
- Sens des mots : désignation du concept lexical que la forme peut exprimer
- Organisée par des relations sémantiques (pointeurs vers des synsets) ou relations lexicales :
  - Relation entre les significations des mots: hyperonymie(inv. Hyponymie) et Méronymie (sous partie de, toit est un méronyme de maison, inv. Holonyme)
  - Relation entre les formes des mots: synonymies, Antonymie (riche/pauvre)

WordNet / Exemple de noms

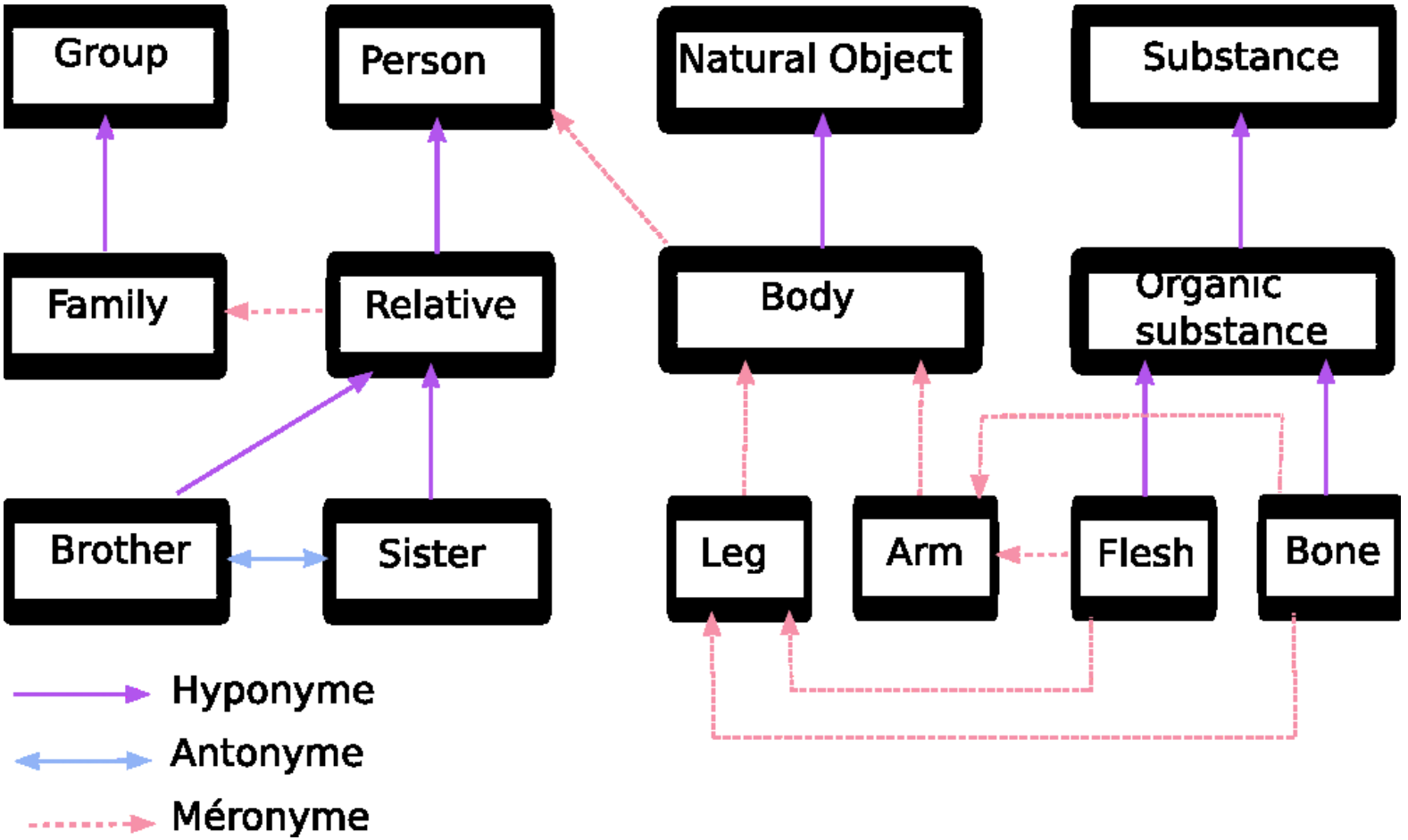


Figure: Exemple de wordnet

Organisation des verbes

- la plus importante catégorie lexicale et syntaxique du langage et la plus difficile à étudier
- 21000 verbes et approximativement 8400 synsets
- répartition des verbes dans 15 domaines (verbes de l'émotion, de la possession, du changement, de la compétition, du mouvement, ...)

Exemple d'utilisation

- Détecter la liste des auteurs dans des documents
- Plusieurs formes très différentes comme **created by ...** ou encore **authors: ...**
- **created by**: **by** supprimé (car mot outil), **created** $\rightarrow$ **create**
- **authors**: lien synomymique vers **creator** puis **create**

Et les autres langues ?

- EuroWordNet pour les langues européennes avec liens vers l'anglais (WordNet). Ressources payantes et couverture moins importante que l'anglais
- EuroWordNet Francais: 22745 synsets, 32809 mots
- Listes des langues/projets : http://www.globalwordnet.org/gwa/wordnet_table.htm

Définition Ontologie

- Ontologie: désigne un ensemble structuré de savoirs dans un domaine de connaissance particulier. Deux parties :
  - la première définit la nature des éléments qui composent le domaine
  - la seconde partie explicite les relations entre plusieurs instances définies dans la première partie
- Vision identique aux réseaux sémantiques mais appliqué au web (en pratique plutôt utilisé dans le domaine de la veille)
- Normalisation via des standards W3C (normaliser le contenu est une démarche beaucoup plus complexe)

RDF

- Standard W3C (2004) en XML
- RDF: permet de définir la connaissance sous forme de triplet (sujet, prédicat, objet).
- Exemple RDF: X auteur de http://www.epita.fr

```
<rdf:Description about="http://www.epita.fr">
  <auteur>X</auteur>
</rdf:Description>
```

- Exemple d'utilisation de RDF: FoAF (Friend of a Friend)

OWL

- Aussi un standard W3C (2004) : extension de RDF
- fait pour représenter des ontologies
- ajout du concept de :
  - classe d'équivalence
  - propritéte d'équivalence
  - égalité/différence de deux ressources
  - de symétrie et de cardinalité

Conclusion

Transformation d'un document en descripteurs

- supprimer les mots outils (stop words)
- étiquetage morpho-syntaxique ou stemming pour les mots
- garder uniquement certaines catégories
- prendre des N-gram sur ces traitements pour éviter les ambiguïtés
- utiliser un dictionnaire sémantique (comme wordnet ou une ontologie du domaine) pour avoir des descripteurs moins sensibles au bruit et des meilleurs similarités entre documents

Bibliographie

1. **Foundations of Statistical Natural Language Processing**, Chris Manning and Hinrich Schütze, MIT Press.
2. **Text Mining: Predictive Methods For Analyzing Unstructured Information**, Sholom M. Weiss, Nitin Indurkhya et Tong Zhang
3. **Social Networks and the Semantic Web**, Peter Mika