L'étiqueteur de parties de discours
Un outil pour le traitement des langues naturelles

Atelier sur les corpus spécialisés en terminologie
28 avril 2003

Denis L'Homme

L'étiqueteur de parties de discours
Un outil pour le traitement des langues naturelles

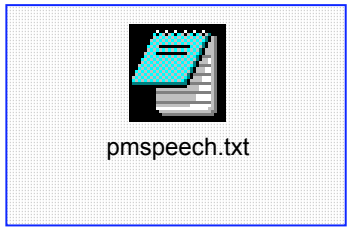
- Plan de présentation
  - 1 - Introduction
  - 2 - Différentes approches à l'étiquetage de textes
    - Manuelle
    - Informatiques
      - Règles codées
      - Basées sur corpus
  - 3 - Zoom sur l'étiquetage basé sur corpus pré étiquetés
  - 4 - L'Étiqueteur à règles apprises d'Eric Brill
    - La méthode d'apprentissage
  - 5 - L'étiquetage
  - 6 - Quel étiqueteur choisir?
  - 7 - Pour en savoir et en avoir plus...

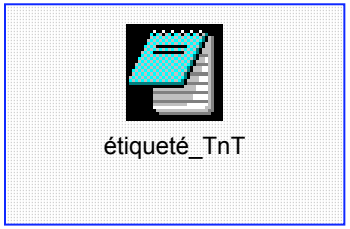
L'étiqueteur de parties de discours
Un outil pour le traitement des langues naturelles

- **Rappel**
  - **L'étiquetage consiste à attribuer à chaque mot d'un texte une étiquette qui représente la partie de discours (nom, adjectif, verbe, etc.) basée sur le contexte dans lequel apparaît le mot**
  - **L'étiquetage fait donc appel à deux sources d'information:**
    - **l'information lexicale: les diverses étiquettes admises pour chaque mot (lexique)**
    - **l'information contextuelle: l'étiquette appropriée compte tenu du contexte (règles)**

L'étiqueteur de parties de discours
Un outil pour le traitement des langues naturelles

- Un texte non étiqueté
- Exemples de textes étiquetés





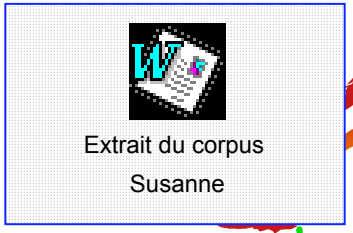
L'étiqueteur de parties de discours
2 - Différentes approches à l'étiquetage de textes

- 1 - Approche manuelle
  - Nombre et types d'étiquettes
  - Règles

Jeu d'étiquettes

- Exemples de corpus étiquetés
  - Brown
    - 500 textes de 2000+ mots dans 15 catégories
  - Wall Street Journal
    - 1 200 000+ mots (articles de journaux)
  - Frantext
    - 180 M mots (textes des 19è et 20è siècles)
  - Susanne
    - Sous-ensemble de Brown 64 textes parmi les 500)
  - ...

Extrait d'un corpus étiqueté Susanne



159 ou 62 étiquettes

L'étiqueteur de parties de discours
2 - Différentes approches à l'étiquetage de textes

- 2- Approches informatiques
  - 1 - Étiquetage basé sur des règles codées (comme pour l'étiquetage manuel)
    - KTAGGER (www.sil.org/computing/catalog/ktagger.html)
  - 2 - Étiquetage basé sur des corpus pré étiquetés (algorithmes d'apprentissage)
    - probabiliste (Markovien)
      - TnT TAGGER (n-gram tagger) (www.coli.uni-sb.de/~thorsten/tnt/)
      - JMX TAGGER (entropie maximale) (www.cis.upenn.edu/~adwait)
    - plus proche voisin
      - QTAG (www.clg.bham.ac.uk/QTAG)
    - règles apprises
      - Brill (www.cs.jhu/~brill/code.html ou http://jupiter.inalf.cnrs.fr)
  - 3- Étiquetage basé sur des corpus non pré étiquetés (algorithmes d'apprentissage)
    - probabiliste (Hidden Markov Models)
      - TLBTAGGER (ftp://lands.let.kun.nl/pub/tosca/tlbtag/)
    - règles apprises
      - étiqueteur à apprentissage non supervisé de Brill
  - 4 - Autres
    - Arbres de décision
    - «Connectionist machines»

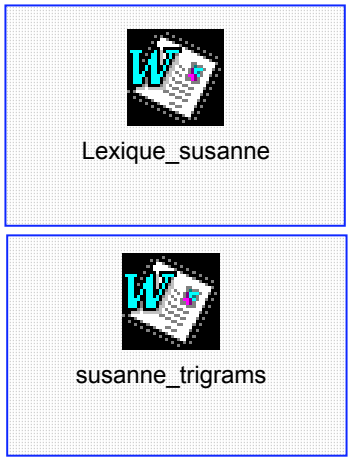
L'étiqueteur de parties de discours
3 - Zoom sur les étiqueteurs basés sur corpus

- D'où provient l'intérêt pour l'étiquetage basé sur corpus?
  - Les méthodes d'analyse linguistique basées sur corpus (dont l'étiquetage) n'ont pas à faire appel aux complexités du langage
  - Ces méthodes tablent sur le fait que des phénomènes linguistiques complexes peuvent être indirectement observés au travers d'épiphénomènes simples tels, pour l'étiquetage, la fréquence et l'ordre des mots
  - Les étiqueteurs basés sur corpus sont portables, c-à-d. qu'ils sont entraînables sur de nouveaux corpus ou jeux d'étiquettes
  - Les ordinateurs d'aujourd'hui rendent possibles, dans un temps raisonnable, l'analyse de volumineux corpus et l'utilisation d'algorithmes d'apprentissage performants

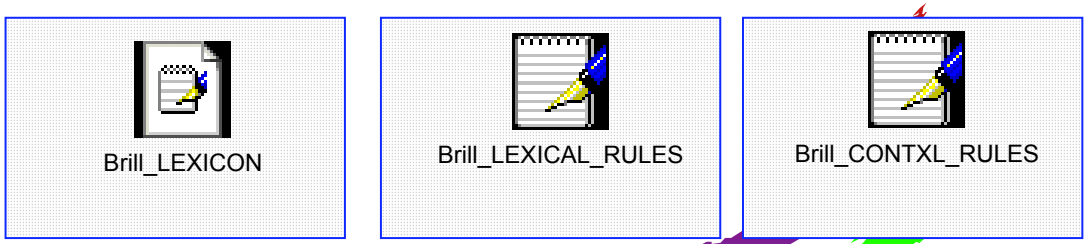
L'étiqueteur de parties de discours
3 - Zoom sur les étiqueteurs basés sur corpus - l'apprentissage

- **À partir d'un corpus étiqueté, l'algorithme d'apprentissage génère les deux types d'information requis pour permettre l'étiquetage d'un texte**
  - **un lexique**
  - **des «règles»**

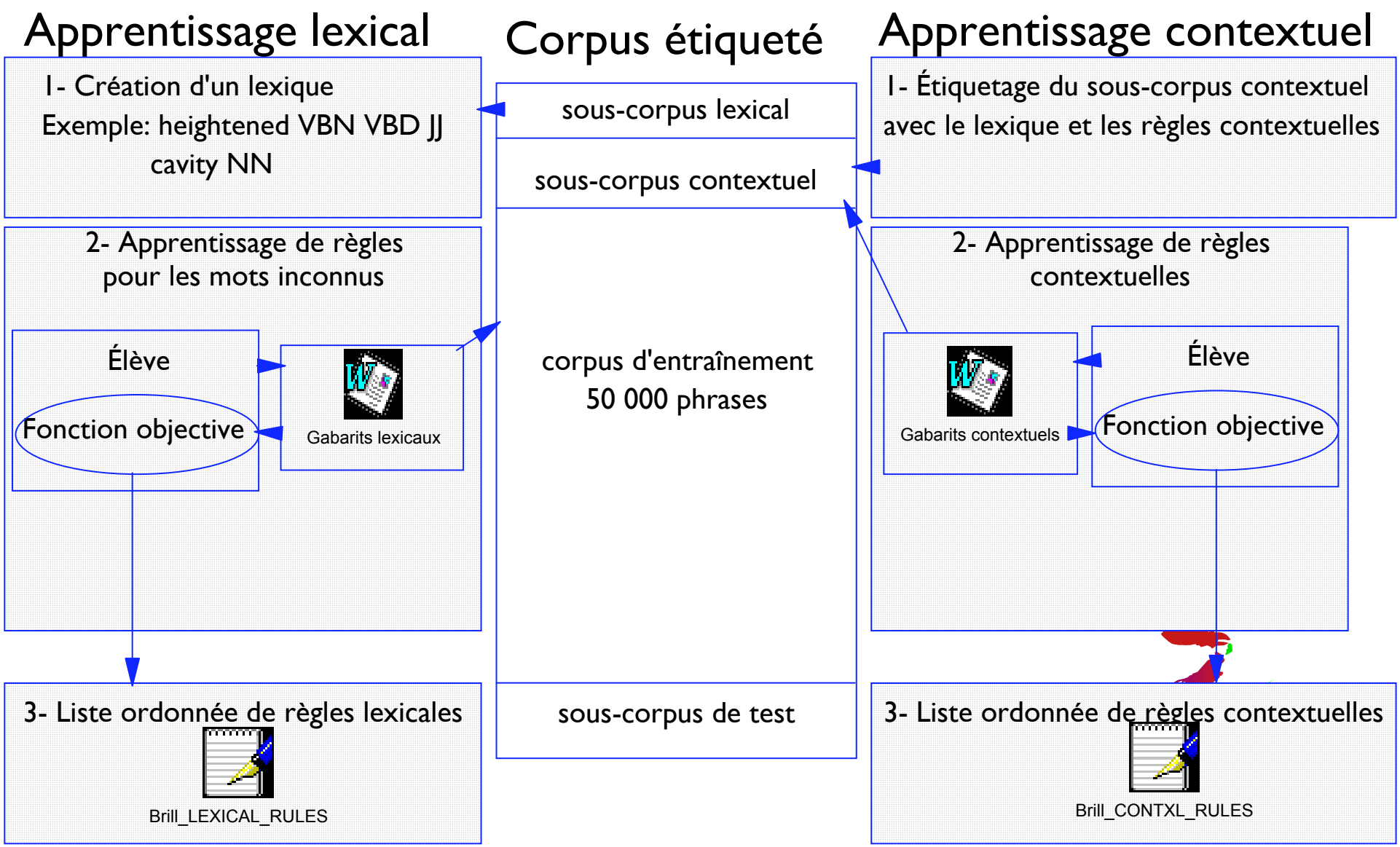
Fichiers générés par TnT



Fichiers générés par Brill



L'étiqueteur de parties de discours
4 - L'étiqueteur à règles apprises de Brill - l'apprentissage



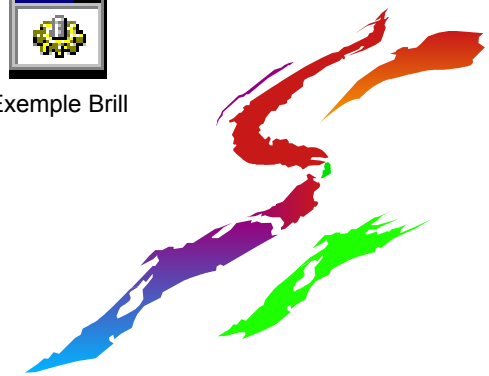
L'étiqueteur de parties de discours
5 - L'étiquetage

- **Avec TnT**
  - **1- Une fois le système «entraîné» on étiquette du nouveau texte en assignant la chaîne d'étiquettes qui maximise P(W|T) x ($P(T_i|T_{i-1}...T_{i-n})$) pour une phrase donnée**
  - **2- Les préfixes/suffixes sont utilisés pour traiter les mots inconnus**

Exemple TnT

- **Avec Brill**
  - **1- Le système attribue d'abord aux mots inconnus l'étiquette NNP ou NN**
  - **2- Il leur applique ensuite les règles lexicales et modifie les étiquettes en conséquence**
  - **3- Il attribue aux mots connus la première étiquette du lexique**
  - **4- Il applique ensuite les règles contextuelles à l'ensemble des mots pour améliorer l'étiquetage global**

Exemple Brill



L'étiqueteur de parties de discours
6 - Quel étiqueteur choisir?

- Pour étiqueter
  - Brill et TnT se valent
    - précision des résultats - 96% à 98 %
    - temps d'exécution (léger avantage TnT)
    - facilité d'utilisation
- ...des textes anglais
  - Brill
    - Brown - 51947 entrées
    - WSJ - 70 697 entrées
    - Brown/WSJ - 93 696 entrées
  - TnT
    - WSJ
    - Susanne - 16 462 entrées
- ...des textes allemands
  - TnT
    - Negra - 58 750* entrées

*estimation

- ...des textes français
  - Brill
    - INAIF - 440 544 entrées
- ...textes français en incluant le lemme
  - Brill version INAIF (Winbrill)

texte_et_lemme

- Pour entraîner avec un nouveau corpus
  - Temps d'exécution
    - Gros avantage TnT
  - Précision des résultats
    - Avantage Brill pour petits corpus d'entraînement
    - Avantage TnT pour gros corpus d'entraînement

L'étiqueteur de parties de discours
6 - Mais à quoi ça sert?

- C'est aux linguistes de le dire
  - Une idée...
    - Un concordancier plus sélectif

L'étiqueteur de parties de discours
7 - Pour en savoir et en avoir plus...

www.cis.upenn.edu/~adwait/penntools.html#/Packages