SL0720Y
Étiquetage morpho-syntaxique
Correction de la séance 1 (5 mars 2009)

Franck Sajous - CLLE-ERSS
http://w3.erss.univ-tlse2.fr/membre/fsajous/

1 À la découverte de Treetagger : interface en ligne

De quelles informations dispose-t-on en plus du texte initial ?
- des partie du discours (*i.e.* catégorie grammaticale, notées ci-après POS, pour *part of speech*) ;
- des lemmes.

2 Installation

3 Utilisation

3.1 tag-french

Quelles sont toutes les fonctions effectuées par ce programme ?
segmentation en mots + segmentation en phrases en plus de l'étiquetage morpho-syntaxique et de la lemmatisation.

3.2 Le programme tree-tagger

`tree-tagger C:\SL0720\TreeTagger\lib\french.par` *monfichier.txt*

Que constatez-vous ?
Sortie = une seule ligne : `ABR`
Le programme `treetagger` attend un texte segmenté en entrée : un token (mot ou signe de ponctuation) par ligne. Si on lui fournit un texte non segmenté, il considère tous les mots d'une même ligne comme un seul token.

Éditez le fichier texte en insérant des retour-chariots. Que constatez-vous ?
Note : *retour-chariot* = retour à la ligne (origine : les machines à écrire).

1. Par défaut, la sortie ne contient que les POS et n'affiche ni les formes, ni les lemmes.
2. La segmentation en mots n'est pas triviale : clitiques, apostrophes et autres. Doit-on accoler le tiret au premier mot, au deuxième ou le considérer comme un token à part entière ?

berce - le	NOM PUN DET:art	berce - le	c ' est	VER:con PUN VER:pres	<unknown> ' être
berce- le	NOM DET:art	<unknown> le	c' est	PRO:dem VER:pres	ce être
berce -le	VER:subp DET:art	bercer le	c 'est	VER:con NOM	<unknown> <unknown>
berce-le	DET:art	<unknown>	c'est	NOM	<unknown>

Note : berce → nf Plante vivace, de la famille des ombellifères

Important : lors de la segmentation manuelle en tokens, une ligne doit contenir un mot ou signe de ponctuation par ligne **sans espace**. Un espace avant ou après un mot est considé comme faisant partie du mot et conduit à une erreur d'étiquetage.

Syntaxe ?

```
treetagger chemin_fichier_param -token -lemma fichier_txt
```

avec `fichier_txt` déjà segmenté. Dans votre cas :

```
treetagger c:\SL0720\TreeTagger\lib\french.par -token -lemma fichier_txt
```

Le script `tag-french` traite un fichier texte non segmenté. Il se charge d'appeler le programme `tokenize.pl` (qui se situe dans `TreeTagger\cmd`) qui effectue automatiquement la segmentation en mots. Le script `tag-french` passe les options `-token` (affichage de la forme) et `-lemma` (affichage du lemme) au programme `treetagger`.
Sauf en cas de nécessité d'utiliser d'autres options, ce script peut être utilisé par la suite pour traiter les textes français.

4 Jeu d'étiquettes

4.1 Pour le français

1. étiquettez ce fichier (forme, partie du discours, lemme).
   (a) qu'observe-t-on dans la colonne des lemmes ?
   `unknown` pour les mots inconnus
   `@card@` pour les nombres
   articles partitifs : du→ de le
   (b) quelles sont les étiquettes pour les marques de ponctuation ?
   `SENT, PUN:cit, PUN`
   (c) à quoi correspondent-elles ?
   fin de phrases (ponctuation fortes), citations (guillemets) et autres.

4.2 Pour l'anglais

Quelle remarque concernant les marques de ponctuation peut-on faire par rapport au français ?
Ponctuations fortes : étiquette = `SENT` (pareil qu'en français)
Autres : POS = forme = lemme

Quelles autres remarques peut-on faire concernant le jeu d'étiquettes ?
Il n'y a pas d'équivalence stricte entre les différentes étiquettes. Certaines catégories sont plus détaillées dans une langue que dans l'autre. Cela peut relever de différences intrinsèques aux langues et des corpus disponibles utilisés pour réaliser l'apprentissage des catégories.

4.3 Cordial

Quelles informations `Cordial` apporte-t-il en plus de `TreeTagger` ?

Les colonnes de la sortie Cordial donnée en figure 2 de l'énoncé correspondent à :

1. numéro de mot dans la phrase
2. forme
3. lemme
4. pos (catégorie grammaticale)
5. pos2 (autre codage pour la catégorie)
6. numéro de la tête du syntagme minimal [| du syntagme maximal]
7. fonction syntaxique
8. numéro de la proposition
9. pivot (verbe de la proposition)
10. type de proposition (eg. S→contient le sujet, V→verbe de base de la proposition, B→appartient à l'attribut du sujet, T→contient le sujet)
11. indications sémantiques (recours à des ontologies)

Quelles autres différences observez-vous ?
Découpage en phrases : 6 phrases avec TT, 20 avec Cordial !

Erreurs d'étiquetage

- *bouche ouverte, tête nue* (vers 5)
  TT : nue → NOM/nue (erreur)
  Cordial : tête nue→ ADV/tête nue (??)
- *étendu dans l'herbe sous la nue* (vers 7)
  TT : nue → NOM/nue (OK)
  Cordial : nue → NCFS/nue (OK)
  Note : *nue* → nuage, ensemble de nuages. P. méton. Le ciel, l'atmosphère, nuageuse ou non.
- *Sourirait un enfant malade, il fait un somme* (vers 10)
  Cordial : "fait un somme" → sourire/VINDP3S
  (ce n'est pas une erreur d'analyse, c'est un bug !)

5 Étiquetage des mots inconnus

5.1 Études de cas

Quand un token présent dans le lexique peut avoir plusieurs POS/lemmes, TT peut trouver le moyen de désambiguïser (cf fig. 1). À ce stade, on peut supposer qu'il s'appuie sur le contexte syntaxique (séquences d'étiquettes plus probables que d'autres). Pour la première ambiguïté, par exemple, dans *cet été*, on peut avoir *été* → NOM/été ou *été* → VConj/être. La présence du token précédent *PRO:DEM* permet de trancher. En effet, PRO:DEM NOM est possible alors que PRO:DEM VConj paraît improbable.

```
Cet     PRO:DEM         ce
été     NOM             été
a       VER:pres        avoir
été     VER:pper        être
le      DET:ART         le
plus    ADV             plus
bel     ADJ             beau
été     VER:pper        être
qu'     KON             que
il      PRO:PER         il
ait     VER:subp        avoir
été     VER:pper        être
depuis  PRP             depuis
l'      DET:ART         le
été     NOM             été
2000    NUM             @card@
.       SENT            .
```

FIG. 1 – *Ce été a été le plus bel été qu'il ait été depuis l'été 2000.*

```
ragnaneur       NOM     <unknown>
```

FIG. 2 – *ragnaneur*

```
ragnanir        VER:infi        <unknown>
```

FIG. 3 – *ragnanir*

```
Un              DET:ART un
grand           ADJ     grand
ragnanir        NOM     <unknown>
```

FIG. 4 – *Un grand ragnanir*

Lorsqu'un token est absent de son lexique, TT peut se baser sur les suffixes (cf fig. 2 et 3). En dépit du suffixe qui inciterait à catégoriser *ragnanir* comme un infinitif, il est étiqueté `NOM` lorsqu'il est précédé de la séquence `DET ADJ` (cf fig. 4).

guerre civile	NOM ADJ	guerre civil
guerre cccivile	NOM ADJ	guerre <unknown>
guerre civilation	NOM NOM	guerre <unknown>

Le recours aux suffixes peut primer sur le contexte syntaxique.

5.2 Probabilités

Treetagger propose l'affichage des différentes étiquettes possibles pour un token :

```
tree-tagger fichier_param -token -lemma
                    -prob -threshold <p> fichier_txt
```

affiche pour un token toutes les possibilités d'étiquettes dont la probabilité est spérieure à `p`.
Exemple :

```
tree-tagger C:\SL0720\TreeTagger\lib\french.par
            -token -lemma -prob -threshold 0.001 fichier_txt
```

un	DET:ART un 0.979 NUM un 0.0159
grand	ADJ grand 1.000
ragnavient	VER:pres <unknown> 0.695 NOM <unknown> 0.289 VER:subp <unknown> 0.008 VER:futu <unknown> 0.003 VER:impf <unknown> 0.001

5.3 Sorties Cordial

Dans le cas de mots connus, Cordial commet une erreur sur le troisème *été* (correctement étiqueté par Treetagger).
Pour les mots inconnus, on peut simplement constater que l'étiquette attribuée à *ragnanir* (nom commun inconnu) ne change pas dans des contextes syntaxiques différents et que malgré le suffixe *-ation*, *civilation* est étiqueté `Adj`.