3 - Etiquetage morpho-syntaxique

1. Principe
2. Etiqueteur n-gramme
3. Etiqueteur à base de règles de correction

Références : *livre en ligne sur NLTK* *http://nltk.org/book/ch05.html*

Jurafsky & Martin - Speech and Language Processing - chapitre 8

3.1 - Principe

- Il s'agit d'annoter un texte déjà tokenisé par des étiquettes représentant une information morpho-syntaxique (on considère que les tokens sont des mots).
- Le but est d'aider au traitement ultérieur en réduisant l'ambiguité lexicale.
- Cette tâche utilise un jeu d'étiquettes qui représentent des parties du discours souvent raffinées par une information morphologique.
- Le choix du jeu d'étiquettes est crucial. S'il est trop fin, la désambiguisation sera efficace mais elle sera plus coûteuse et le risque de commettre des erreurs plus grand. S'il est trop grossier, c'est le contraire.
- Pour évaluer un étiqueteur, on l'applique à un corpus de référence (gold standard) déjà étiqueté et validé par des linguistes et on mesure les erreurs.

3.1 - Principe

Jeu d'étiquettes pour le français de TreeTagger

ABR Abreviation
ADJ Adjectif
ADV Adverbe
DET:ART Article
DET:POS Pronom Possessif (ma, ta, ...)
INT Interjection
KON Conjunction
NAM Nom Propre
NOM Nom
NUM Numéral
PRO Pronom
PRO:DEM Pronom Démonstratif
PRO:IND Pronom Indéfini
PRO:PER Pronom Personnel
PRO:POS Pronom Possessif (mien, tien, ...)
PRO:REL Pronom Relatif

PRP Préposition
PRP:DET Préposition + Article (au,du,aux,des)
PUN Ponctuation
PUN:CIT Ponctuation de citation
SENT Balise de phrase
SYM Symbole
VER:COND Verbe au conditionnel
VER:FUTU Verbe au futur
VER:IMPE Verbe à l'impératif
VER:IMPF Verbe à l'imparfait
VER:INFI Verbe à infinitif
VER:PPER Verbe au participe passé
VER:PPRE Verbe au participe présent
VER:PRES Verbe au présent
VER:SIMP Verbe au passé simple
VER:SUBI Verbe à l'imparfait du subjonctif
VER:SUBP Verbe au présent du subjonctif

3.2 - Etiqueteur n-gramme

- Un étiqueteur n-gramme est fondé sur le principe que l'étiquette d'un mot dépend des étiquettes des n-1 mots qui le précèdent.

- Cette dépendance est exprimée sous forme de probabilités qui sont apprises sur un corpus déjà étiqueté et validé, appelé corpus d'apprentissage.

- Pour que ces probabilités aient un sens, il est nécessaire de disposer d'un corpus d'apprentissage suffisamment gros. Plus n est grand, plus le calcul des probabilités est coûteux et moins les résultats sont fiables, du fait de la dispersion des données.

- Dans la pratique, on se limite à des faibles valeurs de n : 1 (unigramme), 2 (bigramme), 3 (trigramme).

3.2 - Etiqueteur n-gramme

- L'espace probabilisé considéré est formé de phrases étiquetées notées (w_{1n}, e_{1n}) , où w_{1n} est une phrase formée de la suite de mots w_1, ..., w_n et e_{1n} est une suite correspondante d'étiquettes e_1, ..., e_n .

- Etant donné une phrase w_{1n} , il s'agit de trouver la suite d'étiquettes e_{1n} pour laquelle la probabilité $P(e_{1n} \mid w_{1n})$ est maximum.

- Pour un étiqueteur unigramme, l'étiquetage de chaque mot est indépendant de l'étiquetage des mots précédents et suivants, ce qui s'exprime par :

$$P(e_{1n}|w_{1n}) = \Pi_{i=1}^{i=n} P(e_i|w_i)$$

- Chaque probabilité $P(e_i|w_i)$ est calculée sur le corpus d'apprentissage comme le rapport entre le nombres de mots w_i étiquetés e_i au nombre de mots w_i présents dans le corpus. On choisit l'étiquette e_i pour laquelle $P(e_i|w_i)$ est maximum.

3.2 - Etiqueteur n-gramme

- Pour un étiqueteur n-gramme où n est supérieur à 1, la formule de Bayes permet d'inverser la probabilité conditionnelle :

$$P(e_{1n}|w_{1n}) = \frac{P(w_{1n}|e_{1n}) \times P(e_{1n})}{P(w_{1n})}$$

- Maximiser $P(e_{1n}|w_{1n})$ pour une phrase w_{1n} donnée, revient à maximiser :

$$P(w_{1n}|e_{1n}) \times P(e_{1n})$$

- En considérant que le probabilité d'apparition d'un mot wi pour une étiquette donnée ne dépend que de cette étiquette, cette formule se simplifie en :

$$\Pi_{i=1}^{i=n} P(w_i|e_i) \times \Pi_{i=2}^{i=n} P(e_i|e_{1i-1}) \times P(e_1)$$

3.2 - Etiqueteur n-gramme

- Pour un étiqueteur bigramme, on suppose que la probabilité d'apparition d'une étiquette ne dépend que de l'étiquette précédente, et la formule se simplifie en :

$$\Pi_{i=1}^{i=n} P(w_i|e_i) \times \Pi_{i=2}^{i=n} P(e_i|e_{i-1}) \times P(e_1)$$

- Chaque probabilité $P(w_i|e_i)$ d'apparition du mot w_i connaissant son étiquette e_i est calculée sur le corpus d'apprentissage comme le rapport entre le nombre d'occurrences du mot w_i étiqueté e_i par le nombre d'occurrences de e_i.

- Chaque probabilité $P(e_i|e_{i-1})$ d'apparition d'apparition d'une étiquette e_i immédiatement après une étiquette e_{i-1} (bigramme), est calculée sur le corpus d'apprentissage comme le rapport entre le nombre d'occurrences d'étiquettes e_i suivant immédiatement une étiquette e_{i-1} par le nombre d'occurrences de e_{i-1}.

- La probabilité $P(e_1)$ pour une étiquette e_1 d'apparaître en début de phrase est calculée sur le corpus d'apprentissage comme le rapport entre le nombre d'occurrences d'étiquettes e_1 apparaissant en début de phrase par le nombre de phrases.

3.2 - Etiqueteur n-gramme

- Soit les deux phrases à étiqueter avec le TreeTagger:

  Jean achète une ferme

  Jean la ferme

- Le corpus d'apprentissage fournit les résultats suivants :
  - ➡ *Jean* et *achète* ont une seule étiquette possible, respectivement *NAM* et *VER:PRES*
  - ➡ Le mot *une* a deux étiquettes possibles avec les probabilités d'apparition : *P(une|DET:ART) = 0,08, P(une | PRO:IND) = 0,002*
  - ➡ Le mot *la* a deux étiquettes possibles avec les probabilités d'apparition : *P(la|DET:ART) = 0,12, P(la | PRO:PERS) = 0,14*
  - ➡ Le mot *ferme* a trois étiquettes possibles avec les probabilités d'apparition : *P(ferme|NOM) = 0,006, P(ferme|VER:PRES) = 0,04, P(ferme|ADJ) = 0,12*
  - ➡ Pour les bigrammes utiles, voici les probabilités : *P(ADJ | DET:ART) = 0,24, P(ADJ | PRO:IND) = 0,07, P(ADJ | PRO:PERS) = 0,0001, P(DET:ART | NAM) = 0,06, P(DET:ART | VER:PRES) = 0,45, P(NOM | DET:ART) = 0,58, P(NOM | PRO:IND) = 0,0001, P(NOM | PRO:PERS) = 0,00004, P(PRO:IND | VER:PRES) = 0,09, P(PRO:PERS | NAM) = 0,08, P(VER:PRES | DET:ART) = 0,00003, P(VER:PRES | PRO:IND) = 0,005, P(VER:PRES | PRO:PERS) = 0,11,*

3.2 - Etiqueteur n-gramme

- Calcul de l'étiquetage le plus probable pour *Jean achète une ferme*:
  - ➡ *P(une | DET:ART) x P(ferme | NOM) x P(DET:ART | VER:PRES) x P(NOM | DET:ART) = 0,08 x 0,006 x 0,45 x 0,58 = 1,25280 x 10^{-4}*
  - ➡ *P(une | DET:ART) x P(ferme | VER:PRES) x P(DET:ART | VER:PRES) x P(VER:PRES| DET:ART) = 0,08 x 0,04 x 0,45 x 0,00003 = 4,32 x 10^{-8}*
  - ➡ *P(une | DET:ART) x P(ferme | ADJ) x P(DET:ART | VER:PRES) x P(ADJ | DET:ART) = 0,08 x 0,12 x 0,45 x 0,24 = 1,0368 x 10^{-3}*
  - ➡ *P(une | PRO:IND) x P(ferme | NOM) x P(PRO:IND | VER:PRES) x P(NOM | PRO:IND) = 0,002 x 0,006 x 0,09 x 0,0001 = 1,08 x 10^{-10}*
  - ➡ *P(une | PRO:IND) x P(ferme | VER:PRES) x P(PRO:IND | VER:PRES) x P(VER:PRES | PRO:IND) = 0,002 x 0,04 x 0,09 x 0,005 = 3,6 x 10^{-8}*
  - ➡ *P(une | PRO:IND) x P(ferme | ADJ) x P(PRO:IND | VER:PRES) x P(ADJ | PRO:IND) = 0,002 x 0,12 x 0,09 x 0,07 = 1,512 x 10^{-6}*

L'étiquetage le plus probable est donc :

Jean/NAM achète/VER:PRES une/DET:ART ferme/ADJ

3.2 - Etiqueteur n-gramme

- Calcul de l'étiquetage le plus probable pour *Jean la ferme*:
  - ➡ *P(la | DET:ART) x P(ferme | NOM) x P(DET:ART | NAM) x P(NOM | DET:ART) = 0,12 x 0,006 x 0,06 x 0,58 = 2,5056 x 10^{-5}*
  - ➡ *P(la | DET:ART) x P(ferme | VER:PRES) x P(DET:ART | NAM) x P(VER:PRES | DET:ART) = 0,12 x 0,04 x 0,06 x 0,00003 = 8,64 x 10^{-9}*
  - ➡ *P(la | DET:ART) x P(ferme | ADJ) x P(DET:ART | NAM) x P(ADJ | DET:ART) = 0,12 x 0,12 x 0,06 x 0,24 = 2,0736 x 10^{-4}*
  - ➡ *P(la | PRO:PERS) x P(ferme | NOM) x P(PRO:PERS | NAM) x P(NOM | PRO:PERS) = 0,14 x 0,006 x 0,08 x 0,00004 = 2,688 x 10^{-9}*
  - ➡ *P(la | PRO:PERS) x P(ferme | VER:PRES) x P(PRO:PERS | NAM) x P(VER:PRES | PRO:PERS) = 0,14 x 0,04 x 0,08 x 0,11 = 4,928 x 10^{-5}*
  - ➡ *P(la | PRO:PERS) x P(ferme | ADJ) x P(PRO:PERS | NAM) x P(ADJ | PRO:PERS) = 0,14 x 0,12 x 0,08 x 0,0001 = 1,344 x 10^{-7}*

L'étiquetage le plus probable est donc :

Jean/NAM la/PRO:PERS ferme/VER:PRES

3.2 - Etiqueteur n-gramme

- Le modèle des étiqueteurs n-grammes est une application particulière des **chaînes de Markov cachées** (HMM) où on cherche à déterminer la suite de phénomènes cachés qui a la probabilité la plus grande compte tenu d'une suite correspondante d'observations.

- L'algorithme qui est en général utilisé pour le calcul de cette probabilité est l'**algorithme de Viterbi**.

- Du fait de la dispersion des données, certaines suites d'étiquettes ne sont pas présentes dans le corpus d'apprentissage. Pour pallier ce défaut, on combine un étiquetage n-gramme avec les étiquetages n-1-gramme ... jusqu'à unigramme.

- Lorsque des mots ne sont pas présents dans le corpus d'apprentissage, une façon de résoudre le problème est de limiter l'étiquetage du corpus d'apprentissage aux n mots les plus fréquents et de marquer tous les autres avec une nouvelle étiquette INC (inconnu).

3.2 - Etiqueteur n-gramme

1. Soit le corpus suivant :

Un garçon de la ferme entre et pose de la viande sur la table. La fille se tient entre la porte et et le fauteuil de la mère. Elle prend la viande que le garçon a posée et la porte à la cave. D'un ton ferme, le garçon lui dit de prendre garde à la porte car elle ferme mal.

a) Etiquetez ce corpus manuellement en utilisant le jeu d'étiquettes du TreeTagger.

b) Dans la perspective d'utiliser un étiqueteur unigramme avec ce corpus comme corpus d'apprentissage, estimez à partir ce ce dernier les probabilités P(DET:ART | la) , P(PRO.PER | la) , P(VER:PRES | porte), P(PRO.REL | que), P(ADJ | ferme).

c) Dans la perspective d'utiliser un étiqueteur bigramme avec ce corpus comme corpus d'apprentissage, estimez à partir ce ce dernier les probabilités P(DET:ART | VER:PRES), P(NOM | DET.ART), P(DET.ART | NOM), P(PRO.PER | NOM), P(VER.PRES| PRO.PER), P(NOM |PRO.PER), P(ADJ | PRO.PER), P(ADJ | DET.ART), P(la | DET.ART), P(la | PRO.PER), P(ferme | NOM), P(ferme | VER.PRES), P(ferme | ADJ).

d) A partir des calculs précédents, appliquez un étiqueteur bigramme entraîné sur le corpus initial à la phrase : *un garçon la ferme.*

e) On garde seulement les étiquettes des 10 mots les plus fréquents et les autres sont remplacées par INC. Dans ces conditions, appliquez un étiqueteur bigramme entraîné sur le corpus avec ce nouvel étiquetage à la phrase : *un garçon la regarde.*

3.3 - Etiqueteur à base de règles de correction

- Le principe, conçu par Brill, est de partir d'un étiquetage grossier du corpus pour le corriger en lui appliquant une suite de règles de correction.

- Une règle de correction a la forme CG E CD → E' qui a la signification suivante : si on rencontre dans le corpus l'étiquette E et si ses contextes immédiats d'étiquettes gauches et droites sont respectivement CG et CL, alors on remplace E par E'.

- La suite de règles est apprise sur un corpus étiqueté de référence.

3.3 - Etiqueteur à base de règles de correction

- On étiquette le corpus de référence C à l'aide d'un étiqueteur unigramme. On obtient un corpus étiqueté C_1. On cherche ensuite la règle de correction R_1 qui permet de corriger un maximum d'erreurs dans C_1 en utilisant un minimum de contexte. On applique R_1 à C_1 et on obtient un corpus C_2. On réitère l'opération à partir de C_2 jusqu'à ce que le nombre d'erreurs tombe au-dessous d'un certain seuil. On obtient ainsi une suite de règles de correction R_1, R_2, ..., R_m.

- On étiquette le corpus d'application à l'aide d'un étiqueteur unigramme puis on corrige l'étiquetage obtenu à l'aide de la suite R_1, R_2, ..., R_m.

3.3 - Etiqueteur à base de règles de correction

2. Soit le corpus suivant :

Un garçon de la ferme entre et pose de la viande sur la table. La fille se tient entre la porte et et le fauteuil de la mère. Elle prend la viande que le garçon a posée et la porte à la cave. D'un ton ferme, le garçon lui dit de prendre garde à la porte car elle ferme mal.

a) Etiquetez ce corpus manuellement en utilisant le jeu d'étiquettes du TreeTagger.

b) Appliquez ensuite un étiqueteur unigramme à ce corpus. Lorsque deux étiquettes sont en compétition pour un mot, on choisit l'étiquette la plus fréquente dans le corpus.

c) On se sert de ce corpus pour apprendre une suite de règles de correction. Déterminez selon l'algorithme vu en cours, une suite de règles qui permette de corriger le corpus étiqueté avec l'étiqueteur unigramme et qui nécessite le moins d'appel au contexte.