idl_09 : Modèles de Markov cachés et annotation en parties du discours

Introduction

Morphologie en linguistique

- domaine qui traite de la structure interne des mots
- linguistique structurale :
  - morphème = unité linguistique minimale (non décomposable) porteuse de sens
    - unités abstraites
    - immangeable = im- mange -able
  - notion de morphe = (une) forme graphique d'un morphème
    - allomorphes : variantes d'un même morphème
      /al-/ dans allons, /v-/ dans vais, /ir-/ dans irai sont trois allomorphes du même morphème (ici, pour "aller")
    - libres : le/beau peuvent exister seuls
    - liés : -cevoir dans recevoir, percevoir, décevoir, etc.
    - contextuelles : j'/je

Procédés morphologiques

- flexion : déclinaison, conjugaison

  grand/grands/grande, cours/courir/courons

- dérivation : formation de nouveaux mots notamment par adjonction d'affixes au radical

  anti-constitu-tionn-el-le-ment

- composition : combinaison de plusieurs bases pour former un nouveau mot

  tournevis

Étiquetage morpho-syntaxique

Présentation

- Analyse morphologique : segmentation
  but : découper un mot en segments morphémiques
- Analyse morpho-syntaxique
  but : analyser chaque mot pour lui associer divers types d'informations telles que la catégorie grammaticale , des traits morphologiques ainsi que le lemme correspondant, c'est à dire la forme canonique du mot (masculin singulier, infinitif, etc.)

⚠️ Les mots ont généralement plus d'une étiquette possible

- Le bois vient de
  France. -> le/DET bois/NC ...
- Je le bois.
  -> Je le/PRONOM bois/VERBE

Objectif de l'étiquetage: déterminer l'étiquette pour une instance d'un mot

Exemple d'entrée à annoter : Le débat est relancé

- ambiguïtés:
  - le = det / pro
  - débat = verbe / nom
  - est = verbe / nom
- Exemple de sortie:
  - Le/DET débat/V est/NOUN relancé/V
    📖 regardons la probabilité de cette séquence d'étiquettes
    Grewmatch(https://universal.grew.fr/?corpus=UD_French-GSD@2.13#)

```
pattern { X1 [upos=DET]; X2 [upos=VERB]; X3 [upos=VERB]; X4
[upos=VERB]; X1 < X2; X2 < X3; X3 < X4 }
```

- sortie correcte :
  - Le/DET débat/NC est/V relancé/V

- Applications:
  - synthèse vocale: comment prononcer "couvent" ?
  - recherche dans un corpus: "est" en tant que nom
  - entrée d'un analyseur syntaxique

> 📖 recherche sur Grewmatch (https://universal.grew.fr/?corpus=UD_French-GSD@2.13#)
>
> ```
> pattern { X [upos="NOUN", form="est"] }
> ```

❓ Comment choisir l'étiquette adaptée ?

- contexte :

Le	bois	vient	de	France
DET	NC	V	P	NPP
PRO	V	V	P	NPP

- connaissance des probabilités d'étiquettes des mots
  la bonne soupe😋 🍜 ou la bonne soupe👩 🍴 ?

La	bonne	soupe
DET	NC (p=0.01)	V (p=0.1)
DET	ADJ (p=0.95)	NC (p=0.85)

Exercice

> Orange, au cours du matin, a permis la prise de bénéfices. Nous avions fait des paris audacieux sur cette valeur qui est montée dès l'ouverture du CO.

Etiquettes possibles : adjectif, adverbe, déterminant, interjection, verbe, participe passé, nom commun, nom propre, préposition, préposition+déterminant, pronom.

- Orange [nom propre, adjectif, nom]
- au [préposition+déterminant]
- cours [nom, verbe]
- du [préposition+déterminant, participe passé, nom (mal orthographié ?)]

- au cours du [préposition]
- matin [adverbe, nom]
- a [verbe, nom]
- permis [participe passé, nom, adjectif]
- la [déterminant, pronom, nom]
- prise [nom, participe passé]
- de [préposition]
- bénéfices [nom]

Lemmatisation :

- Orange -> orange | Orange
- au -> à + le
- cours -> cours | courir
- du -> de + le | du | devoir
- matin -> matin
- a -> a | avoir
- permis -> permis | permettre
- la -> le
- prise -> prise | prendre
- de -> de
- bénéfices -> bénéfice

Quelques approches

Morpho-syntaxe : solutions (?) en extension

Le lexique Morphalou (https://www.ortolang.fr/market/lexicons/morphalou) 3.1 comprend la transcription phonétique de 93 695 lemmes et 522 662 formes fléchies.
La version 3 de Morphalou (https://www.ortolang.fr/market/lexicons/morphalou) a été obtenue par la fusion de quatre lexiques :

- Morphalou 2 (version de décembre 2013)
- DELA (version de décembre 2011)
- Dicollecte (version 4.3)
- LGLex et LGLexLefff (version 3.4)
- Lefff (version 2.1 avril 2006)

"Factoriser" la lemmatisation : la racinisation

Une solution naïve énumérative.

1. recenser les affixes
    - grammaticaux (flexionnels) : ons, ont, s, aux
    - morphèmes : -ment, -eur, in-
2. tronquer, à droite ou à gauche
    - part/sort/ -ir -ie -ant -ons
    - malad/ -e -es -ie -ies

=> construction de graphes : Démonette (https://demonette.fr/demonext/vues/front_page.php)
exemple du lemme "pas" (https://demonette.fr/demonext/vues/lexeme_family_graph.php?lid=l316928)

❓ Peut-on facilement développer un modèle d'annotation en morphosyntaxe (catégories grammaticales) facilement avec la bonne ressource ?

Modèles de Markov cachés (Hidden Markov Model, HMM)

Les modèles de Markov cachés sont des modèles de Markov où les états du systèmes ne sont pas les observables (= cachés). L'objectif ici est d'apprendre à deviner les états sous-jacents étant donné des données observables.

- observables : séquence de mots
- états cachés : catégories grammaticales

Situation initiale :

Les catégories grammaticales forment une chaîne de Markov.
Dans un HMM, une phrase est une séquence d'étiquettes morphosyntaxiques ayant servi à générer une séquence de mots.

Modélisation d'une phrase annotée en morphosyntaxe dans un HMM :

Le modèle

Un HMM utilise trois types de probabilités :

- les probabilités initiales (flèche épaisse) : les probabilités de chaque étiquette d'apparaître en début de séquence.
- les probabilités de transition (flèches vers la droite) : les probabilités de passer d'une étiquette à une autre (= chaîne de Markov d'ordre 1).
- les probabilités d'émission (flèches vers le bas) : les probabilités qu'un état (étiquette) émette une observation (mot).

Construction depuis exemples

Se fait à l'aide de statistiques, comme pour une chaîne de Markov.

- la/DET bonne/NC soupe/VERB
- la/DET bonne/ADJ soupe/NC
- la/DET soupe/NC fume/VERB

Initiales :

- DET = 1

Transitions :

- ADJ , NC = 1
- DET , NC = 2/3
- DET , ADJ = 1/3
- NC , VERB = 1

Émissions :

- ADJ -> bonne = 1
- DET -> la = 1
- NC ->
  - bonne =1/3
  - soupe =2/3
- VERB ->
  - soupe =1/2
  - fume =1/2

exemple repris de : https://lattice.cnrs.fr/sites/itellier/cours-HMM-CRF.pdf (https://lattice.cnrs.fr/sites/itellier/cours-HMM-CRF.pdf)

Décodage : test de notre modèle sur la phrase "la bonne fume"

On appelle, pour un HMM, le décodage le fait de trouver la meilleure séquence d'états étant donnée une séquence d'obervations données.

L'algorithme de Viterbi utilise la propriété de Markov pour donner la meilleure séquence. À chaque instant, on évalue quel est le meilleur "pas" à faire.

La meilleure séquence est simplement la suite des meilleurs pas (la séquence la plus probable est la suite des transitions les plus probables).

L'algorithme fonctionne de la façon suivante :

1. pour le 1er mot : on calcule, pour chaque étiquette, sa probabilité initiale étant donné le mot.
2. pour les autres : on calcule, pour chaque étiquette, les meilleures transitions depuis le mot précédent, on garde une trace de là où l'on vient (back pointer).
3. récupération la meilleure séquence en partant de la fin et en rebroussant chemin (backtrack).

Algorithme de Viterbi

On commence avec une matrice vide (penser que cellule vide = probabilité de 0). On calcule la probabilité du premier mot en multipliant la probabilité initiale et la probabilité d'émission.

- Initiales : DET = 1
- DET -> la = 1

	la	*bonne*	*fume*
ADJ	0		
DET	1		
NC	0		
VERB	0		

Pour chaque étiquette à un instant t (étiquette_t) possible, on cherche la meilleure transition depuis l'instant $t-1$

Transitions :

- ADJ , NC = 1
- DET ,
  - NC = $2/3$
  - ADJ = $1/3$
- NC , VERB = 1

On calcule la meilleure transition pour chaque étiquette du mot suivant : "bonne".

	la	*bonne*	*fume*
ADJ	0		
DET	1		
NC	0		
VERB	0		

	la	*bonne*	*fume*
ADJ	0	$\mathbf{1} \times \frac{1}{3}$	
DET	**1**		
NC	0		
VERB	0		

Une fois la meilleure transition trouvée, on la multiplie avec la probabilité d'émission.

Émissions :

- ADJ -> bonne = 1
- DET -> la = 1
- NC ->
  - bonne =1/3
  - soupe =2/3
- VERB ->
  - soupe =1/2
  - fume =1/2

	la	*bonne*	*fume*
ADJ	0	$\frac{1}{3} \times 1$	
DET	1		
NC	0		
VERB	0		

	la	*bonne*	*fume*
ADJ	0	$\frac{1}{3}$	
DET	1	0	
NC	0		
VERB	0		

- DET ,
  - NC = 2/3
  - ADJ = 1/3

	la	*bonne*	*fume*
ADJ	0	$\frac{1}{3}$	
DET	**1**	0	
NC	0	$\mathbf{1} \times \frac{2}{3}$	
VERB	0		

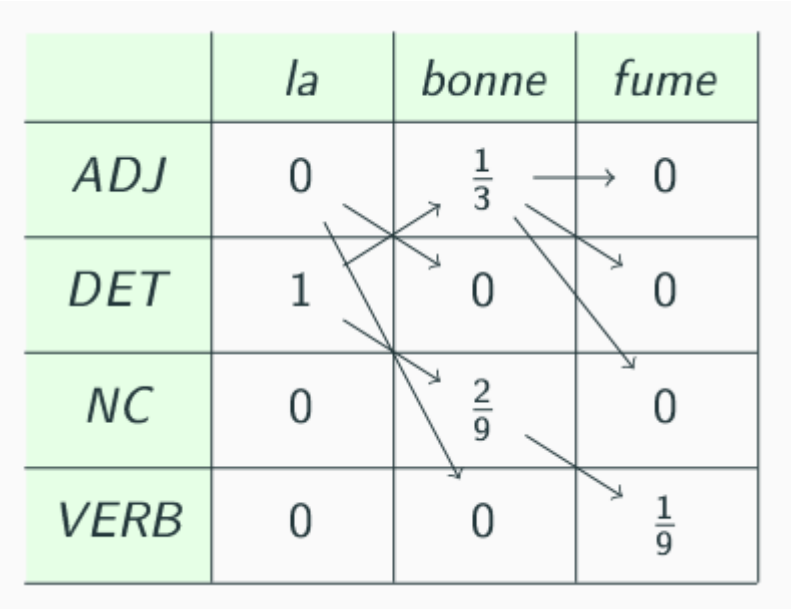
- NC ->
  - bonne = 1/3

	la	*bonne*	*fume*
ADJ	0	$\frac{1}{3}$	
DET	1	0	
NC	0	$\frac{2}{3} \times \frac{1}{3}$	
VERB	0		

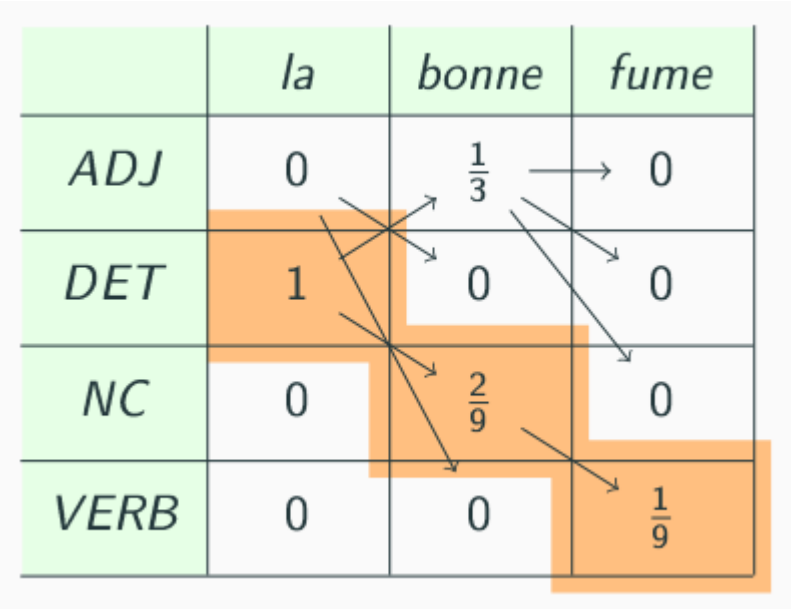
	la	*bonne*	*fume*
ADJ	0	$\frac{1}{3}$	
DET	1	0	
NC	0	$\frac{2}{9}$	
VERB	0	0	

... On continue à remplir la matrice :

- NC , VERB = 1
- VERB ->
  - soupe =1/2
  - fume =1/2



... Puis on backtrack pour trouver la séquence la plus probable !



Résultat : "La bonne fume" => la/DET bonne/NC fume/VERB

Quelques ressources

- ELDA (European Language Ressource Association) http://www.elda.org/ (http://www.elda.org/)
- CNRTL (Centre National de Ressources Textuelles et Lexicales) http://www.cnrtl.fr/ (http://www.cnrtl.fr/)
- Delaf http://infolingu.univ-mlv.fr/ (http://infolingu.univ-mlv.fr/)
- GeoNames http://www.geonames.org/ (http://www.geonames.org/) étude pour les technologies du langage.
- LeFFF (Lexique des Formes Fléchies du Français)

http://alpage.inria.fr/~sagot/ (http://alpage.inria.fr/~sagot/)

- Morphalou https://www.ortolang.fr/market/lexicons/morphalou (https://www.ortolang.fr/market/lexicons/morphalou)
- les lexiques ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue) https://www.ortolang.fr/market/lexicons (https://www.ortolang.fr/market/lexicons)

NB : les ressources sont à la fois un besoin, une finalité et un objet