Lexique et morphologie

18 novembre 2019

—Kata Gábor
kata.gabor@inalco.fr—

Tokens, formes et lemmes

Je le dis maintenant, comme je le disais hier, que le prix du diesel est élevé, mais que le prix de l'essence est encore plus élevé.

- **tokens** ou occurrences : 27 (31 avec ponctuations) : tous les mots courants ou mots + signes de ponctuation
  - tokenisation : séparer le texte en tokens, identifier les tokens qui contiennent des ponctuations et/ou des espaces
  - l'essence, USA/U.S.A./U. S. A., 06 71 75 98 06, H_2O
  - normalisation : *l'essence* → *l ' essence, don't* → *do not, Je* → *je ...*
- **formes/types** : le (4), élevé (2), que (2), prix (2), est (2), Je, je, comme, de, ...
- **lemmes ou lexèmes** : je, être, dire, comme, de, diesel, ...

Morphologie computationnelle

- **analyse morphologique** : étant donné une forme, y associer la catégorie grammaticale, les catégories flexionnelles (et le lemme)

	lemme	cat. gr + flex
dit	dire	V +sg +3 + pres
	dire	V +partpass
	dire	V +sg +3 +psimple

- **génération** morphologique : étant donné un lemme et des catégories flexionnelles, trouver la forme correspondante parler, V + sg + 1 + imparf → parlais
- **tâches connexes** : tokenisation, normalisation, reconnaissance des entités nommées (avant) ; lemmatisation (au lieu de), étiquetage morphologique ou POS tagging (après)
- applications : correcteurs d'orthographe, moteurs de recherche, traduction automatique : analyse et génération, systèmes d'interaction homme-machine, reconnaissance de la parole

Analyse morphologique automatique

L'analyse morphologique consiste à définir une relation M (pour Morphologie) sur deux ensembles :

1. l'ensemble F des formes de mots (aller, va, ira, allé)
2. l'ensemble C des descriptions : lemmes + catégories flexionnelles (aller+V+inf, aller+V+sg+3)

$$M \subseteq F \times C \tag{1}$$

où

$$F \times C = \{(f, c) | f \in F \ \ et \ \ c \in C\} \tag{2}$$

Paradigme

— **paradigme d'une catégorie grammaticale** : *toutes* les catégorie flexionnelles pour une partie du discours (p.ex. verbe : temps passé, présent, futur, mode indicatif, subjonctif, ...)
— **paradigme d'un mot** : toutes les formes associées à un lemme

puella (*fille*)		
Cas	Sg	Pl
Nom	puella	puellae
Acc	puellam	puellas
Gen	puellae	puellarum
Dat	puellae	puellis
Abl	puella	puellis

Approches de description linguistique

1. énumération mot + paradigme (word and paradigm)
2. concaténation de morphèmes, radical + affixes
3. approche procédurale (règles de production de formes ; une règle par catégorie flexionnelle)

Stratégies I. L'approche "lookup"

- stocker *toutes* les formes avec les analyses correspondantes dans un dictionnaire
- lire l'entrée et rechercher dans le dictionnaire
- contre : ignore les régularités dans la morphologie
- pré-requis : *générer* toutes les formes (!)
  - largement possible pour l'anglais, faisable pour le français
  - turc, arabe, hébreu, allemand, hongrois, russe : très problématique
  - autre problème avec le lookup : productivité et nouveaux mots dans le corpus

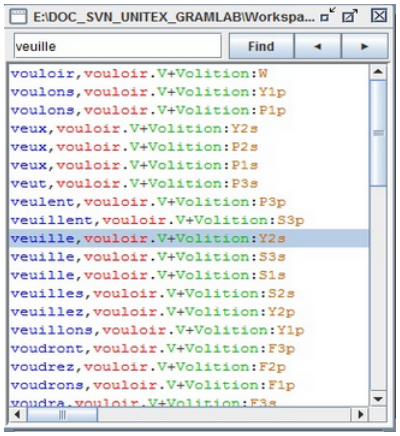


FIGURE 1 – Dictionnaire pour la méthode lookup : formes et analyses

- analyse avec méthode lookup : automate fini déterministe acyclique (arbre de recherche, *search tree*, voir plus tard)

II. Analyse avec régularités

Outil : automates et transducteurs à états finis

Un **automate à états finis** est constitué de :

- V, un vocabulaire ou alphabet **fini**
- Q, un nombre fini d'états dont
  - état initial (un ou plusieurs)
  - état final (un ou plusieurs)
- f, fonction de transition : donne pour chaque état $q \in Q$ et chaque élément du vocabulaire $v \in V$, le ou les états que l'on peut atteindre en partant de q et en utilisant v : $f(q,v) \subset Q$.

Un **transducteur à états finis** contient en plus :

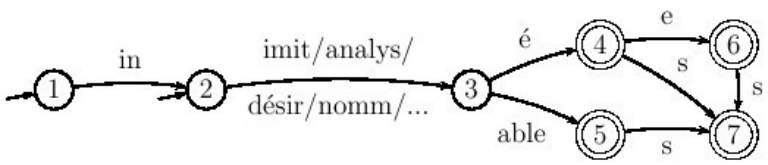
— un vocabulaire ou alphabet de sortie S,
— une fonction de sortie s qui associe à un état q et une entrée v une sortie s $\in$ S

P.ex. : automate permettant d'analyser les mots comme : *"inimitables", "imitée", "analysées", "inanalysable", "désirable", "indésiré", "innomable"*

V = {in, imit, désir, analys, nomm, able, é, e, s}

— racines *imit, désir, analys, nomm*
— préfixe *in*
— suffixe *able, é, e, s*

visualisation de l'automate par graphe :



— ici, les états sont représentés par des noeuds numérotés 1-7
— états initiaux : ceux sur lesquels arrive une flèche qui vient de nulle part : ici, 1 et 2
— états finaux : 4, 5, 6, 7 (double rond par convention)
— transitions : flèches/liens entre les états avec étiquette de transition (élément du vocabulaire), p ex f(3,able)=5
— les boucles (transition d'un état vers lui-même) sont en principe autorisées

Un **chemin** dans un automate commence nécessairement à partir d'un état initial, suit des transitions et aboutit à un état final.

Le **langage** accepté/défini par l'automate est l'ensemble de concaténations d'éléments du vocabulaire V qui correspondent à un chemin.

Il existe d'autres automates reconnaissant le même langage. Par exemple, un automate avec seulement un état initial et un état final, et autant de transitions que de mots distincts conviendrait tout aussi bien. Mais celui de la figure met en évidence des régularités de construction que n'auraient pas l'autre (équivalent à une simple liste).

Avantages des automates :

- bien adaptés pour modéliser l'affixation :
  - conjugaison en français
  - déclinaisons dans les langues à cas
  - dérivations productives
- rapidité
- facile à combiner par des opérations d'union, concaténation, ...

Les automates finis reconnaissent une classe de langages particulière appelée langages réguliers ou langages rationnels.

Exercices

1. Construisez un automate morphologique qui reconnaît

   - la conjugaison des verbes français en -er : à l'indicatif présent, imparfait, futur

   - les déclinaisons latines suivantes :

Cas	Singulier	Pluriel
Nominatif	Rosa	Rosae
Accusatif	Rosam	Rosas
Génitif	Rosae	Rosarum
Datif	Rosae	Rosis
Ablatif	Rosa	Rosis

Cas	Singulier	Pluriel
Nominatif	Murus	Muri
Accusatif	Murum	Muros
Génitif	Muri	Murorum
Datif	Muro	Muris
Ablatif	Muro	Muris

   - Transformez les automates en transducteurs capables d'associer une analyse aux formes reconnues.