Acquisition de la structure morphologique du lexique basée sur la similarité lexicale et l'analogie formelle

Nabil Hathout
Nabil.Hathout@univ-tlse2.fr

Université de Toulouse
CLLE-ERSS, CNRS & UTM

UE M2R TAL – 21 octobre 2008

Morphologie computationnelle lexématique

Comment réaliser une analyse morphologique sans recourir aux notions de morphème, d'affixe ni d'exposant morphologique ?

Vue d'ensemble

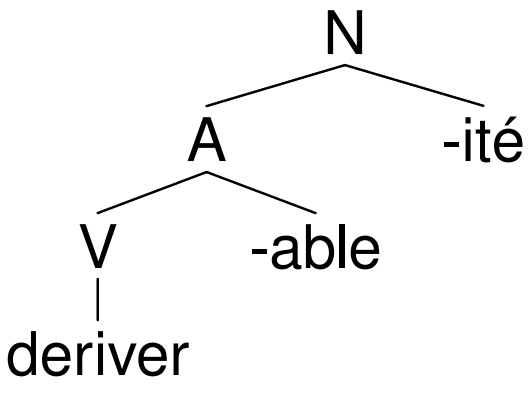
1. Pour une morphologie computationnelle lexématique
   - Morphologie morphématique vs lexématique
   - Associer la similarité morphologique et l'analogie formelle
2. Similarité morphologique
   - Traits formels et sémantiques
   - Connecter les lexèmes à leurs traits
   - Estimer la similarité morphologique entre les mots
   - Voisinages morphologiques
3. Analogies
   - Exploiter les analogies familiales et sérielles
   - Analogies formelles
   - Mise en œuvre
4. Premiers résultats
5. Conclusion

Université de Toulouse

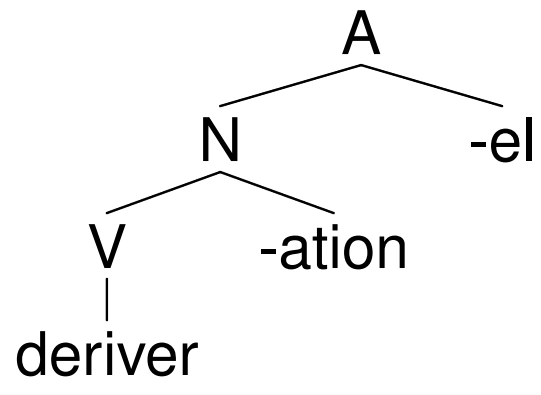
Morphologie morphématique standard

- Les mots sont composés de morphèmes.
- Les morphèmes se composent relativement à des règles de flexion, de dérivation et de composition.
- La structure morphologique des mots est habituellement représentées sous forme arborescente.

dérivabilité



dérivationnel



Problèmes de la morphologie morphématique

- Quel est le statut des morceaux *per-*, *-cevoir*, *-ception* ?

	con-	dé-	per-	re-/ré-
-cevoir	concevoir	décevoir	percevoir	recevoir
-ception	conception	déception	perception	réception

- Pourquoi n'y a-t-il pas de verbe *locomoter* ?

base	dérivé en *-tion*	dérivé en *-if*
dériver_V	dérivation_N	dérivatif_A
?	locomotion_N	locomotif_A

- Il faut un morphème zéro pour analyser les conversions : $[\text{pouvoir}_V\ \emptyset]_N$
- Quel est le statut des interfixes : tarte $\rightarrow$ tart**el**ette ; goutte $\rightarrow$ gutt**el**ette ; vedette $\rightarrow$ vedett**ar**iat

Morphologie lexématique

- Les unités minimales sont les mots. Les mots n'ont pas de structure.
- La structure morphologique est un niveau d'organisation lexical.
- La structure morphologique est composée des relations qui s'établissent entre les mots, notamment :
  - les relations entre les mots de la même famille morphologique, et
  - les relations entre les mots de la même série dérivationnelle.

famille de *dérivation*

dériver
dérivable
dérivatif
dérivationnel
dérivabilité
etc.

série de *dérivation*

acclimatation
compilation
éducation
localisation
variation
etc.

Analyse morphologique

Morphologie morphématique

- découper le mot en une séquence de morphèmes

Morphologie lexématique

- découvrir les relations qui existent entre le mot analysé et les autres unités du lexique
- identifier sa famille morphologique et sa série dérivationnelle

dérivation

mettre en relation *dérivation* avec un nombre suffisant de mots

- de sa famille morphologique : *dériver*, *dérivationnel*, *dérivable*, *dérive*, *dériveur*, etc., et
- de sa série dérivationnelle : *formation*, *séduction*, *variation*, *émission*, etc.

Analogies morphologiques

dérivation et *dérivable* participent à une série d'analogies

dérivation :*dérivable* : :*variation* :*variable*
dérivation :*dérivable* : :*modification* :*modifiable*
dérivation :*dérivable* : :*adaptation* :*adaptable*
dérivation :*dérivable* : :*observation* :*observable*

dérivation et *variation* participent à une série d'analogies

dérivation :*variation* : :*dériver* :*varier*
dérivation :*variation* : :*dérivationnel* :*variationnel*
dérivation :*variation* : :*dérivabilité* :*variabilité*
dérivation :*variation* : :*dérivable* :*variable*

Université de Toulouse

Acquisition des familles morphologiques et des séries dérivationnelles

- Méthode pour l'acquisition des relations morphologiques à partir d'un dictionnaire informatisé *Trésor de la Langue Française informatisé* (TLFi).
- La méthode repose sur
  - une mesure de la similarité morphologique, et
  - la découverte d'analogies formelles entre voisins morphologiques.
- Les 2 techniques sont complémentaires :
  1. les voisinages morphologiques peuvent être calculés aisément pour un grand nombre de mots, mais ils sont trop grossiers pour discriminer entre les mots qui sont effectivement apparentés et ceux qui ne le sont pas ;
  2. les analogies formelles permettent de réaliser un filtrage fin sur les voisins morphologiques, mais sont coûteuses en temps de calcul.

Université de Toulouse

Caractéristiques du modèle

La méthode proposée :

- est purement lexématique et purement relationnelle ;
- intègre de manière uniforme les informations sémantiques et formelles ;
- permet de cumuler des informations de natures différentes ou issues de ressources différentes ;
- rapproche les mots qui partagent le plus grand nombre de traits les plus spécifiques ;
- est compatible avec le modèle « surfaciste » proposé par Burzio (2002) ;
- a une efficacité computationnelle suffisante pour pouvoir être utilisée pour construire des ressources morphologiques semi-automatiquement.

Université de Toulouse

Caractéristiques du modèle

La méthode articule différents travaux :

- la représentation du lexique sous la forme d'un graphe et son exploitation au moyen de parcours aléatoires dans la lignée des travaux de Bruno Gaume
- l'exploitation de la version XML du TLFi (collaboration avec Philippe Muller et Bruno Gaume)
- les travaux de Yves Lepage et de François Yvon et Nicolas Stroppa sur les analogies formelles. La mise en correspondance est réalisée directement sur les représentations graphémiques des mots. Pas de notion de morphèmes.
- l'exploitation des informations sémantiques pour l'acquisition de relations morphologiques (analogies morpho-synonymiques).

Université de Toulouse

Caractéristiques des familles et des séries

- L'appartenance aux familles et aux séries est graduelle

Dans la famille morphologique de *dérive*

dérive		dériveur		dérivationnellement
.	...	.	...	.
0		d_1		d_2

$d_1 < d_2$

Dans la série dérivationnelle de *formation*

formation		compilation		vision
.	...	.	...	.
0		d_1		d_2

$d_1 < d_2$

- Les familles sont des petits ensembles
- Les séries sont des ensembles plus grands
- Les familles ont une cohésion sémantique et formelle plus forte que les séries.

Université de Toulouse

Traits formels et sémantiques

- Deux mots sont morphologiquement reliés s'ils partagent à la fois des propriétés phonologiques et sémantiques.
- Nous utilisons les propriétés graphémiques à la place des propriétés phonologiques parce que le TLFi ne fournit pas la prononciation de toutes les entrées
- la similarité morphologique est estimée en utilisant un bi-graphe qui contient :
  - un ensemble de sommets qui représentent les lexèmes, et
  - un autre ensemble pour leurs propriétés formelles et sémantiques.

Université de Toulouse

Propriétés formelles

- Les traits formels associés à un lexème sont les *n*-grammes de lettres qui apparaissent dans son lemme.
- $n \geq 3$
- Le début et la fin du lemme sont marqués par des `$`.

Les traits formels de *orientation*

`$or;$ori;$orie;$orien;$orient;$orienta;$orientat;`
`$orientati;$orientatio;$orientation;$orientation$;`
`ori;orie;orien;orient;orienta;orientat;orientati;`
`orientatio;orientation;orientation$;... ati;atio;`
`ation;ation$;tio;tion;tion$;ion;ion$;on$`

- Tous les *n*-grammes jouent le même rôle :
  - Ils rapprochent les mots qui contiennent les mêmes sons.
- Aucun *n*-gramme n'a la statut de morphème.

Université de Toulouse

Propriétés formelles

- Les propriétés graphémiques sont réduites à celles des lemmes.
- Il faudrait utiliser comme traits formels d'un lexème les *n*-grammes qui apparaissent dans l'ensemble de ses formes fléchies.
- Avantage : les allomorphies flexionnelles seraient disponibles au niveau dérivationnel
- Inconvénient : les marques flexionnelles réduisent l'homogénéité des traits formels.
  - pour $n \geq 3$, le verbe *malaxer* se retrouve connecté à tous les mots qui contiennent `xie` (*anxieux*, *lexie*, *orthodoxie*, etc.) à cause de la forme *malaxiez*.
  - L'utilisation d'informations sur la fréquence des formes permettrait de réduire l'influence de ce trait formel.

Université de Toulouse

Propriétés sémantiques

- Les traits sémantiques d'une entrée sont les *n*-grammes de mots qui apparaissent dans ses définitions.
- Les *n*-grammes qui contiennent des ponctuations dont ignorés.
- Les définitions sont catégorisées et lemmatisées.

Définition de *orientation*

Action d'orienter, de s'orienter ; résultat de cette action.

Traits sémantiques induits

```
N.action;N.action X.de;N.action X.de V.orienter;
X.de;X.de V.orienter;V.orienter;X.de V.s'orienter;
V.s'orienter;N.résultat;N.résultat X.de;N.résultat
X.de X.ce;N.résultat X.de X.ce N.action;X.de X.ce;
X.de X.ce N.action;X.ce; X.ce N.action;N.action
```

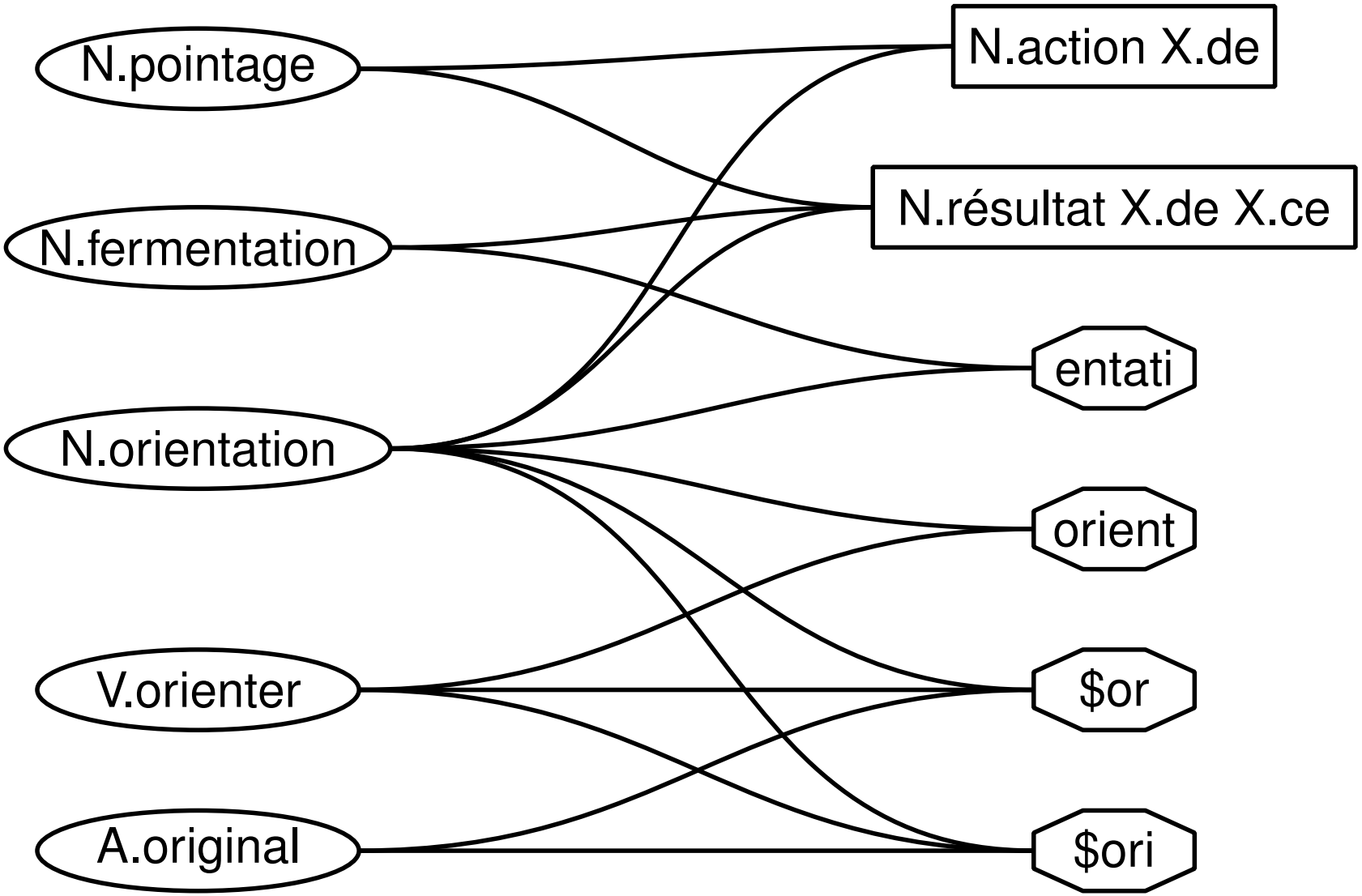
Propriétés sémantiques

- les définitions du TLFi sont « passées à la moulinette »
- représentation très grossière de la sémantique des mots inspirée des segments répétés de Lebart et Salem
  - représentation redondante destinée à capter les ressemblances entre les définitions
  - permet d'intégrer des informations de nature syntagmatique sans réaliser une véritable analyse syntaxique
  - gomme légèrement les variations qui existent dans le traitement des entrées (découpage en sous-sens ; rédaction des définitions)

Un graphe de lexèmes et de traits

- Les traits sémantiques et formels sont utilisés dans le même graphe.
- Le graphe connecte symétriquement chaque entrée à ses traits sémantiques et formels.

Extrait du graphe lexical



Un graphe bi-partie

- La structure de graphe bi-partie n'est pas essentielle : nous cherchons seulement à mesure la similarité morphologique.
- Elle rend possible la propagation synchrone d'une activation dans les sous-graphes formels et sémantiques.
- Le graphe contient la représentation de propriétés qui sont utiles pour la description et l'analyse morphologique (non computationnelle)
- Elles peuvent être utilisées pour :
  - décrire la sémantique de la suffixation en *-able*, ou pour
  - déterminer les finales caractéristiques des noms de bateau : *voilier*, *pétrolier*, *bananier*, *thonier*... *patrouilleur*, *caboteur*, *dériveur*, *dragueur*...

Université de Toulouse

Propagation d'activation

- La similarité morphologique est estimée en propageant une activation dans le graphe.
- L'activation est propagée un nombre pair de fois.
- Dans un graphe fortement redondant 2 étapes de propagation permettent d'obtenir les proximités visées.
- La propagation d'activation est simulée par des parcours aléatoires.
- Elle est calculé en multipliant la matrice d'adjacence stochastique du graphe.

Université de Toulouse

Matrice d'adjacence stochastique

- Soit $G = (V, E)$ un graphe, $V = \{v_1, \cdots, v_n\}$ un ensemble de sommets, $E \subset V \times V$ un ensemble d'arcs.
- La matrice d'adjacence de G est une matrice $n \times n$ telle que :
  $A_{ij} = 1$ si $(v_i, v_j) \in E$
  $A_{ij} = 0$ sinon.
- Les lignes de A sont normalisées : $M_{ij} = \frac{A_{ij}}{\sum_{k=0}^{n} A_{ik}}$
- $(M^n)_{ij}$ est la probabilité d'atteindre v_j en partant de v_i après un parcours de n étapes.
- Cette valuation des arcs prend en compte :
  - le nombre de traits de chaque entrée, et
  - la fréquence des traits.

Matrice d'adjacence stochastique

- Dans l'expérience qui a été réalisée, l'activation est propagée pour moitié vers les traits formels et pour moitié vers les traits sémantiques.
- Les arcs du graphe bi-partie peuvent être répartis en 3 ensembles J, K, L tels que $E = J \cup K \cup L$ où :
  - J contient les arcs qui relient une entrée à un trait formel ;
  - K contient les arcs qui relient une entrée à un trait sémantique ;
  - L contient les arcs qui relient un trait formel ou sémantique à une entrée.
- La valuation de M est définie comme suit :
  - si $e_{ij} = (v_i, v_j) \in J$, $M_{ij} = 0.5 \frac{A_{ij}}{\sum_{e_{ih} \in J} A_{ih}}$ si v_i est connecté à un trait sémantique et $M_{ij} = \frac{A_{ij}}{\sum_{e_{ik} \in J} A_{ik}}$ sinon.
  - si $e_{ik} = (v_i, v_k) \in K$, $M_{ik} = 0.5 \frac{A_{ik}}{\sum_{e_{ih} \in K} A_{ih}}$ si v_i est connecté à un trait formelle et $M_{ik} = \frac{A_{ik}}{\sum_{e_{ih} \in K} A_{ih}}$ sinon.
  - si $e_{il} = (v_i, v_l) \in L$, $M_{il} = \frac{A_{il}}{\sum_{e_{ih} \in L} A_{ih}}$.

Université de Toulouse

Un graphe lexical construit à partir du TLFi

- Le graphe est construit à partir des entrées et des définitions du TLFi
- Les emplois non standards sont supprimés (archaïques, vieux, argotiques, régionalismes, etc.).
- L'extraction et le nettoyage des définitions a été réalisé en collaboration avec Bruno Gaume et Philippe Muller.
- 225 529 définitions ; 75 024 entrées.
- Les traits associés à une seule entrée sont supprimés

Hapax legomena

traits	complet	réduit	hapax
formels	1 306 497	400 915	69%
sémantiques	7 650 490	548 641	93%
total	8 956 987	949 556	90%

Université de Toulouse

40 premiers voisins morphologiques du verbe *fructifier*

Dans un graphe qui contient seulement des :

traits formels

V.fructifier N.fructification A.fructificateur A.fructifiant A.fructifère V.sanctifier V.rectifier A.rectifier V.fructidoriser N.fructidorien N.fructidor **N.fructuosité R.fructueusement A.fructueux** N.rectifieur A.obstructif A.instructif A.destructif A.constructif **N.infructuosité R.infructueusement A.infructueux V.transsubstantifier V.substantifier V.stratifier V.schistifier V.savantifier V.refortifier V.ratifier V.quantifier V.présentifier V.pontifier V.plastifier V.notifier V.nettifier V.mystifier V.mortifier V.justifier V.idiotifier V.identifier**

- Beaucoup de verbes en *-ifier* parmi les voisins

Université de Toulouse

40 premiers voisins morphologiques du verbe *fructifier*

Dans un graphe qui contient seulement des :

traits sémantiques

V.fructifier V.trouver N.missionnaire N.mission A.missionnaire N.saisie N.police N.hangar N.dîme N.ban V.affruiter N.melon N.saisonnement N.azédarach A.fruitier A.bifère V.saisonner N.roman N.troubadour V.contaminer N.conductibilité N.alevinage V.profiter **A.fructifiant** N.pouvoir V.agir N.opération V.placer N.rentabilité N.jouissance N.avocat N.report **A.fructueux** V.tourner V.chiper N.économat N.visa N.société N.réserve N.récréance

- Beaucoup de noms parmi les voisins, mais aucun verbe en *-ifier*.

40 voisins morphologiques du verbe *fructifier*

dans un graphe qui contient à la fois des :

traits sémantiques et formels

V.fructifier A.fructifiant N.fructification A.fructificateur V.trouver **A.fructifère V.rectifier V.sanctifier** A.rectifier V.fructidoriser N.fructidor N.fructidorien N.missionnaire N.mission A.missionnaire **A.fructueux R.fructueusement N.fructuosité** N.rectifieur N.saisie N.police N.hangar N.dîme N.ban A.fruitier V.affruiter A.instructif A.obstructif A.destructif A.constructif N.conductibilité V.saisonner N.melon N.saisonnement N.azédarach A.bifère V.contaminer N.roman N.troubadour N.alevinage

- Mélange entre les voisins formels et sémantiques

Voisinages morphologiques

- Les traits formels sont les plus prédictifs.
- Les traits sémantiques sont les moins fiables.
- Les membres de la famille tendent à apparaître avant ceux de la série.
- Les traits formels et sémantiques partagés par les membres de la même famille sont plus spécifiques que ceux qui sont partagés par les membres de la même série.
- La similarité morphologique n'est pas suffisamment sélective.

Université de Toulouse

Analogies morphologiques

Les membres des familles morphologiques et des séries dérivationnelles sont impliqués dans de très nombreuses analogies

fructifier :*fructification* forment des analogies avec

rectifier :rectification
certifier :certification
plastifier :plastification
sanctifier :sanctification
vitrifier :vitrification
etc.

fructifier :*sanctifier* forment des analogies avec

fructification :sanctification
fructificateur :sanctificateur
fructifiant :sanctifiant

Analogie et voisins morphologiques

- *fructifier* :*fructification* : :*rectifier* :*rectification*
- *fructification* appartient à la famille de *fructifier*.
- *rectifier* appartient à la série de *fructifier*.
  - $\Rightarrow$ *fructification* et *rectifier* sont des voisins morphologiques de *fructifier*.
- *rectification* appartient à la famille de *rectifier* et à la série de*fructification*
  - $\Rightarrow$ *rectification* est un voisin morphologique à la fois de *rectifier* et de *fructification*.
- Les voisinages morphologiques peuvent être directement utilisés pour rechercher des analogies formelles :
  - Pour chaque entrée *a*,
  - si *b* et *c* sont des voisins de *a*, et
  - si *d* est un voisin de *b* et de *c*,
  - alors vérifier si *a* : *b* :: *c* : *d* est une analogie formelle.

Université de Toulouse

Analogie et voisins morphologiques

- Les analogies permettent de filtrer efficacement les voisins morphologiques
- Si v est morphologiquement apparenté à m, alors
- v est un voisin de m.
- v est soit un élément de la famille de m
  soit un élément de la série de m.
- Il existe alors un autre voisin v' de m tel que
  il existe w voisin de v et de v' tel que $m : v :: v' : w$.
- v' appartient à la famille de m si v appartient à la série de m ou vice versa
- On a ainsi deux configurations possibles :
  1. si $v \in F(m)$, alors $\exists v' \in S(m), \exists w \in S(v) \cap F(v'), m : v :: v' : w$
  2. si $v \in S(m)$, alors $\exists v' \in F(m), \exists w \in F(v) \cap S(v'), m : v :: v' : w$

  où $F(x)$ est la famille de x et $S(x)$ la série de x.

Exemple d'analogie formelle

- les analogies formelles sont vérifiées sur les chaînes de caractères
- les différences entre le premier et le second couple de mots doivent être identiques

fructifier :*fructification* : :*rectifier* :*rectification*

fructifi	er	rectifi	er
fructifi	cation	rectifi	cation

Autres exemples

fructeux :*infructueusement* : :*soucieux* :*insoucieusement*

ϵ / in	fructueu / fructueu	x / sement		ϵ / in	soucieu / soucieu	x / sement	

kataba :*maktoubon* : :*fa3ala* :*maf3oulon*

- transcription de l'arabe 'écrire' :'écrit' : :'faire' :'effet'

ϵ / ma	k / k	a / ϵ	t / t	a / ou	b / b	a / on		ϵ / ma	f / f	a / ϵ	3 / 3	a / ou	l / l	a / on			

Factorisation

- Les analogies formelles peuvent être définies en utilisant la notion de factorisation
- Soit L un alphabet, $a \in L^\star$ une chaîne de caractères définie sur L et $n \in \mathbb{N}$
- On appelle factorisation de a de longueur n une séquence de n chaînes de caractères $f_1, \cdots, f_n \in L^{\star n}$ telle que $a = f_1 \oplus \cdots \oplus f_n$ où $\oplus$ représente la concaténation
- (ma, k, ϵ, t, ou, b, on) est une factorisation de longueur 7 de maktoubon.

Factorisation

- Soient $(a, b, c, d) \in L^{\star 4}$ 4 chaînes de caractères.
- $a : b :: c : d$ est une analogie formelle ssi il existe
  - $n \in \mathbb{N}$
  - 4 factorisations de longueur n des 4 chaînes de caractères $(f(a), f(b), f(c), f(d)) \in L^{\star 4}$ telles que
  - $\forall i, 1 \leq i \leq n, (f_i(b), f_i(c)) \in \{(f_i(a), f_i(d)), (f_i(d), f_i(a))\}$

kataba :*maktoubon* : :*fa3ala* :*maf3oulon*

	f_1	f_2	f_3	f_4	f_5	f_6	f_7
a	ϵ	k	a	t	a	b	a
b	ma	k	ϵ	t	ou	b	on
c	ϵ	f	a	3	a	l	a
d	ma	f	ϵ	3	ou	l	on

Comparaison des séquences d'opérations d'édition

- Opérations d'éditions permettant de transformer une chaîne de caractères en une autre :
  - insertion d'un caractère (`abcd` → `abecd`)
  - suppression d'un caractère (`abcd` → `abd`)
  - remplacement d'un caractère (`abcd` → `abed`).
- La séquence des opérations d'édition peut être déduite de la table de distances d'édition de Levenshtein.
- La distance indique le coût des opérations d'éditions permettant de passer d'une chaîne à l'autre.
- Le coût est de 1 pour l'insertion, la suppression et le remplacement d'un caractère par un caractère différent et de 0 pour les caractères identiques.

Université de Toulouse

Tableau de distance d'édition

		fructueux :	*infructueusement*													
	◇	i	n	f	r	u	c	t	u	e	u	s	e	m	e	n	t
◇	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
f	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15
r	2	2	2	3	2	3	4	5	6	7	8	9	10	11	12	13	14
u	3	3	3	3	3	2	3	4	5	6	7	8	9	10	11	12	13
c	4	4	4	4	4	3	2	3	4	5	6	7	8	9	10	11	12
t	5	5	5	5	5	4	3	2	3	4	5	6	7	8	9	10	11
u	6	6	6	6	6	5	4	3	2	3	4	5	6	7	8	9	10
e	7	7	7	7	7	6	5	4	3	2	3	4	5	6	7	8	9
u	8	8	8	8	8	7	6	5	4	3	2	3	4	5	6	7	8
x	9	9	9	9	9	8	7	6	5	4	3	3	4	5	6	7	8

- ◇ représente le début de la chaîne de caractères.
- La case (i, j) du tableau indique la distance entre la sous-chaîne constituée des i premiers caractères de *fructueux* et des j premiers de *infructueusement*.

Séquence d'opérations d'édition de coût minimal

- On extrait de la table la séquence de coût minimal définie comme suit :
- on part de la dernière case du tableau
- on sélectionne pour chaque case, la case voisine de plus faible coût
- en cas d'égalité
  - préférer la case qui se trouve sur la diagonale (identité ou substitution)
  - à défaut la case de droite (insertion),
  - sinon la case du haut (suppression).

Séquence d'opérations d'édition de coût minimal

	◇	i	n	f	r	u	c	t	u	e	u	s	e	m	e	n	t
◇	0	← 1	← 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
f	1	1	2	↖ 2	3	4	5	6	7	8	9	10	11	12	13	14	15
r	2	2	2	3	↖ 2	3	4	5	6	7	8	9	10	11	12	13	14
u	3	3	3	3	3	↖ 2	3	4	5	6	7	8	9	10	11	12	13
c	4	4	4	4	4	3	↖ 2	3	4	5	6	7	8	9	10	11	12
t	5	5	5	5	5	4	3	↖ 2	3	4	5	6	7	8	9	10	11
u	6	6	6	6	6	5	4	3	↖ 2	3	4	5	6	7	8	9	10
e	7	7	7	7	7	6	5	4	3	↖ 2	3	4	5	6	7	8	9
u	8	8	8	8	8	7	6	5	4	3	↖ 2	← 3	← 4	← 5	← 6	← 7	8
x	9	9	9	9	9	8	7	6	5	4	3	3	4	5	6	↖ 7	8

Signature analogique

- la séquence d'opération est décrite comme une correspondance entre deux factorisations

I	I	M	M	M	M	M	M	M	M	I	I	I	I	I	S
ϵ	ϵ	f	r	u	c	t	u	e	u	ϵ	ϵ	ϵ	ϵ	ϵ	x
i	n	f	r	u	c	t	u	e	u	s	e	m	e	n	t

I	I	M	M	M	M	M	M	M	I	I	I	I	I	S
ϵ	ϵ	s	o	u	c	i	e	u	ϵ	ϵ	ϵ	ϵ	ϵ	x
i	n	s	o	u	c	i	e	u	s	e	m	e	n	t

- Les séquences d'insertions (I), de suppressions (D) et de substitutions (S) sont les mêmes pour les deux couples de chaînes de caractères.
- L'analogie formelle n'implique pas la notion de morphème ni celle d'affixe.

Signature analogique

- fusionner les séquences de caractères identiques :

	I	I	M	I	I	I	I	I	S
a	ϵ	ϵ	fructueu	ϵ	ϵ	ϵ	ϵ	ϵ	x
b	i	n	fructueu	s	e	m	e	n	t

	I	I	M	I	I	I	I	I	S
a	ϵ	ϵ	soucieu	ϵ	ϵ	ϵ	ϵ	ϵ	x
b	i	n	soucieu	s	e	m	e	n	t

- ((I,ϵ,i), (I,ϵ,n), (M, fructueu, fructueu), (I,ϵ,s), (I,ϵ,e), (I,ϵ,m), (I,ϵ,e), (I,ϵ,n), (S,x,t))
- ((I,ϵ,i), (I,ϵ,n), (M, soucieu, soucieu), (I,ϵ,s), (I,ϵ,e), (I,ϵ,m), (I,ϵ,e), (I,ϵ,n), (S,x,t))

Signature analogique

- on associe à chaque couple une signature d'édition σ qui consiste à ne pas spécifier les sous-chaînes communes :
- $\sigma(\texttt{fructueux}, \texttt{infructueusement}) = \sigma(\texttt{soucieux}, \texttt{insoucieusement}) =$ ((I,ϵ,i), (I,ϵ,n), (M, @, @), (I,ϵ,s), (I,ϵ,e), (I,ϵ,m), (I,ϵ,e), (I,ϵ,n), (S,x,t))

Vérification des analogies formelles

$a : b :: c : d$ forment une analogie formelle
si $\sigma(a, b) = \sigma(c, d)$ ou bien si $\sigma(a, c) = \sigma(b, d)$.

Résultats préliminaires

- Nous avons calculé les 100 premiers voisins des entrées du TLFi
- Les voisinages ont été utilisés pour constituer des analogies

Exemples d'analogies correctes collectées

A.fructueux :A.affectueux : :S.infructuosité :S.inaffectuosité
A.fructifiant :A.fructificateur : :A.glorifiant :A.glorificateur
A.frugivore :A.végétivore : :R.frugalement :R.végétalement
A.fruitarien :A.végétarien : :S.fruitarisme :S.végétarisme
A.fruitier :A.laitier : :S.fruiterie :S.laiterie
R.fructueusement :R.affectueusement : :S.fructuosité :S.affectuosité
S.fructification :S.identification :V.fructifier :V.identifier

Exemples d'erreurs

* A.fruité :S.fruste : :A.truité :S.truste
* S.fruit :S.frumentaire : :A.instruit :A.instrumentaire
* S.fruiterie :S.friterie : :V.effruiter :V.effriter

Résultats préliminaires

- Nous avons révisé manuellement les analogies de 22 entrées qui appartiennent à 4 familles morphologiques.

configuration	analogies	corrects	erreurs
formel	169	163	3.6%
sémantique	5	5	0.0%
sém + form	130	128	1.5%

- L'utilisation des traits sémantiques améliore la précision mais réduit le nombre total d'analogies.

De meilleurs résultats pour les mots longs

- La qualité des résultats dépend fortement de la longueur des mots.

long.	analog	corrects	erreurs
4	29	14	51.7%
5	22	14	36.4%
6	8	7	12.5%
7	10	8	20.0%
8	55	54	1.8%
9	29	27	6.9%
10	30	30	0.0%
11	32	32	0.0%
12	19	19	0.0%
13	11	11	0.0%
14	35	35	0.0%
15	63	63	0.0%
16	39	39	0.0%

- 13 groupes de 5 mots sélectionnés aléatoirement.
- la méthode s'appuie principalement sur la similarité formelle.
- La similarité formelle est plus forte pour les mots plus longs.

Tâches à venir

- Créer une amorce en utilisant uniquement les mots les plus long, puis appliquer une méthode par bootstrap
- séparer les familles des séries en s'appuyant sur la structure des sous-graphes de voisins qui participent à des analogies
- prendre en compte la fréquence des signatures des analogies et le nombre d'analogies qui incluent chaque couple de lexèmes
- créer un base de données au format CELEX.
  - concevoir un algorithme capable de segmenter les mots en s'appuyant sur les familles et les séries
- réaliser des caractérisations sémantiques de séries dérivationnelles ou de familles morphologiques
- réaliser le même type d'expérience sur l'anglais afin de pouvoir évaluer la méthode proposée d'une manière plus standard.

Université de Toulouse

Discussion

- Les bases sont remplacées par le ou les membres de la familles qui sont sémantiquement les plus proches.
- Les propriétés formelles et sémantiques sont déconnectées, ce qui permettra à terme de traiter sans difficulté les dérivations dites «parasynthétiques» comme cancer :anti-cancereux
- La méthode est essentiellement computationnelle
  - les hypothèses théoriques sont minimales
  - la méthode exploite essentiellement la mémoire et la puissance de calcul des processeurs
- Il faudrait utiliser en complément du dictionnaire, d'autres sources, notamment des caractérisations sémantiques acquises à partir de corpus analysés syntaxiquement

Université de Toulouse