Etiquetage morphosyntaxique avec identification des unités polylexicales

Matthieu Constant

Université Paris-Est Marne-la-Vallée, LIGM

3 décembre 2010

Introduction

Motivation

- Etiquetage morphosyntaxique: assigner à chaque token une catégorie grammaticale
- Identification des unités polylexicales

Exemple

Max mange une pomme de terre.

Introduction

Motivation

- Etiquetage morphosyntaxique: assigner à chaque token une catégorie grammaticale
- Identification des unités polylexicales

Exemple

Max/NPP mange/V une/DET pomme_de_terre/NC ./PONCT

Etat-de-l'art

- Modèles probabilistes discriminants: Champs Markoviens Conditionnels (CRF), Maximum d'entropie (MaxEnt), Séparateurs à vaste marge (SVM), etc.
- Couplage avec des ressources lexicales externes (Denis et Sagot, 2009)
- Evaluation avec tokenisation parfaite (ex. *a priori* $\rightarrow$ *a_priori*)

Notre contribution

- Couplage de ressources lexicales externes avec le modèle CRF
- Identification d'unités polylexicales (pas de tokenisation parfaite!)
- Encore plus de ressources lexicales
- Intégration de machines à états finis

Nos ressources

- Langue: français
- Corpus annoté: Corpus arboré de Paris 7 (FTB)
- Ressources lexicales : dictionnaires de mots simples et composés, grammaires locales

Corpus annoté French TreeBank (FTB) (Abeillé et al. 2003)

- Corpus arboré converti en corpus annoté en parties du discours
- Composition: 569,080 tokens, 29 étiquettes
- Tokens: mots simples et composés, ponctuations, nombres
- Jeu d'étiquettes: combinaison des 13 catégories principales et des 34 sous-catégories (Crabbé et al, 2008)

Extrait du FTB

```
, PONCT
soit CC
une DET
augmentation NC
de P
1_,_2 DET
% NC
par_rapport_au P+D
mois NC
précédent ADJ
```

Découpage du FTB

	TRAIN	DEV	TEST	total
proportion	80%	10%	10%	100%
#tok	455 264	56 908	56 908	569 080
#MWTok	25 662 (5,6%)	3 738 (6,6%)	3 577 (6,3%)	32 977 (5,8%)

- proportion : proportion en tokens
- #tok : nombre de tokens
- #MWTok: nombre de tokens fusionnés, i.e. tokens multi-mots (*par_rapport_au*) et nombres (*1_,_2*)

Jeu d'étiquettes (Crabbé et al, 2008)

ADJ	adjective
ADJWH	interrogative adjective
ADV	adverb
ADVWH	interrogative adverb
CC	coordination conjunction
CLO	object clitic pronoun
CLR	reflexive clitic pronoun
CLS	subject clitic pronoun
CS	subordination conjunction
DET	determiner
DETWH	interrogative determiner
ET	foreign word
I	interjection
NC	common noun
NPP	proper noun

Jeu d'étiquettes (suite)

P	preposition
P+D	preposition+determiner amalgam
P+PRO	prepositon+pronoun amalgam
PONCT	punctuation mark
PREF	prefix
PRO	full pronoun
PROREL	relative pronoun
PROWH	interrogative pronoun
V	indicative or conditional verb form
VIMP	imperative verb form
VINF	infinitive verb form
VPP	past participle
VPR	present participle
VS	subjunctive verb form

Lexiques morphosyntaxiques

DELA (Courtois 1990, Courtois et al. 1997)

- Construction manuelle
- 746 198 formes simples et 272 228 formes composées

Lefff (Sagot 2010)

- Construction semi-automatique
- Extrait automatiquement de l'étiqueteur MElt (Denis et Sagot, 2009)
- 553 140 entrées fléchies dont 26 311 composées

Grammaires locales (Gross 1997)

Types d'unités polylexicales

- Entités nommées: noms d'organisation, noms de personnes, etc. (Martineau et al. 2009)
- Dates (Blanc et al. 2007)
- Prépositions locatives,
- Déterminants numériques (Constant et al. 2003)

Quelques chiffres

- 211 graphes
- 6 989 formes différentes trouvées dans FTB

Couverture des ressources lexicales dans le FTB

- Pourcentage de token-mots du FTB présents dans les ressources lexicales

Lefff	DELA	Lefff+DELA	Lexique
93,4%	95,9%	97,1%	97,5%

- Pourcentage de token-mots inconnus et absents des ressources lexicales

Corpus	Tokens inconnus	Tokens inconnus et absents du Lefff	Tokens inconnus et absents du DELA	Tokens inconnus et absents du lexique
DEV	5,0%	2,5%	1,5%	1,4%
TEST	4,5%	2,3%	1,4%	1,3%

Complémentarité des dictionnaires

- DELA a une plus grande couverture que Lefff
- Lefff a le même jeu d'étiquettes que FTB
- DELA a un jeu d'étiquettes différent de FTB (analyses linguistiques parfois différentes)

Etiquetage morphosyntaxique statistique

Principe

- Soit une séquence de tokens $x = x_1 x_2 \ldots x_n$
- Soit une séquence candidate d'annotations $y = y_1 y_2 \ldots y_n$
- But : trouver la séquence d'annotations y* telle que:

$$y* = argmax_y P(y|x)$$

Exemple

- x = je cours .
- y*=CLS V PONCT
- sélectionné parmi {CLS V PONCT,CLS NC PONCT,...}

Deux phases pour l'étiquetage statistique

Estimation des paramètres du modèle probabiliste

- Réalisée à partir d'un ensemble d'exemples de séquences déjà annotées
- Dans notre cas, corpus d'apprentissage (TRAIN)

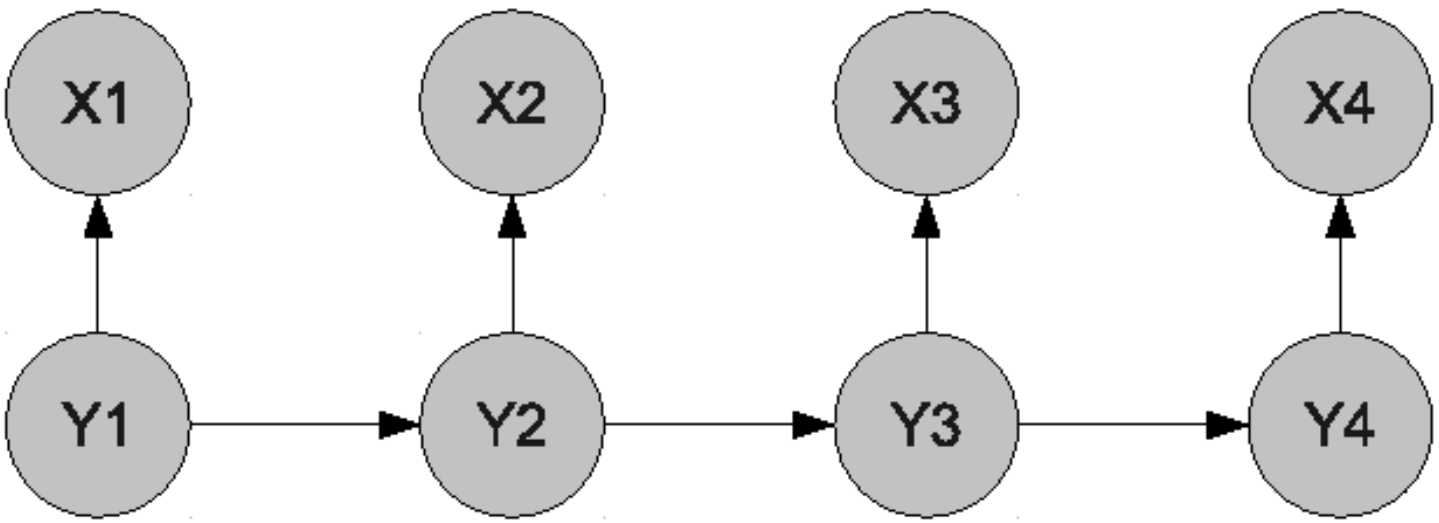
Annotation/décodage

- Etant donné une nouvelle séquence de tokens x, trouver la meilleure séquence y associée
- Programmation dynamique (ex. Viterbi)

Modèle de Markov Caché (HMM)

Dépendances très limitées

- Un token X_i ne dépend que de l'étiquette Y_i indépendamment de la position i
- L'étiquette Y_i ne dépend que de l'étiquette précédente Y_{i-1} indépendamment de la position i



Modèle de Markov Caché (HMM)

Formule d'un HMM du premier ordre

$$P(y|x) = \prod_i P(x_i|y_i).P(y_i|y_{i-1})$$

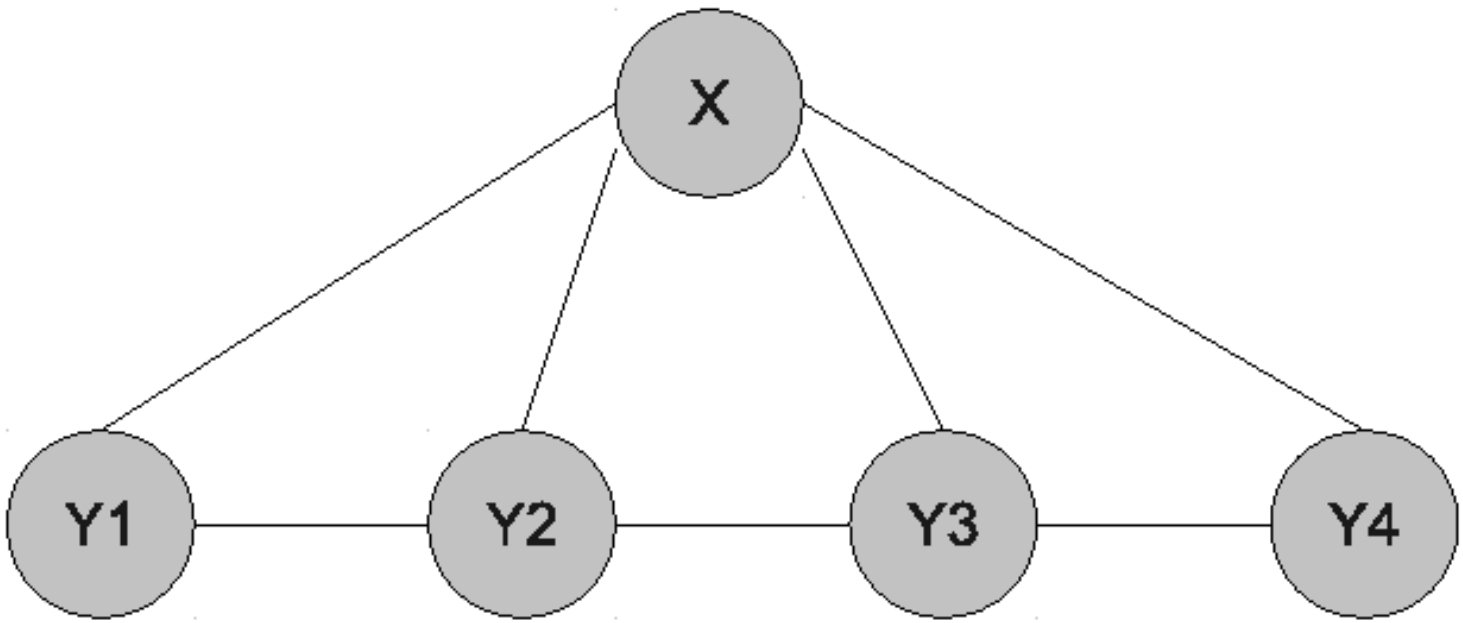
Exemple

P(CLS V PONCT|je cours .) = P(je|CLS).P(CLS|*BOS*)
.P(cours|V).P(V|CLS)
.P(.|PONCT).P(PONCT|V)

Modèle CRF linéaire

Modèle graphique

- Dépendance mutuelle entre X, Y_i, Y_{i-1}



Modèle CRF linéaire

Traits

Les dépendances caractérisées par des traits

- un trait k correspond à une fonction f_k et un poids λ_k
- $f_k(i, y_i, y_{i-1}, x) = 1$ si x,y_i et y_{i-1} ont une certaine configuration à la position i
- $f_k(i, y_i, y_{i-1}, x) = 0$ sinon

Exemples de traits

$$f_{100}(i, y_i, y_{i-1}, x) = \begin{cases} 1 & \text{si } x_i = \text{"le" ET } y_i = \text{"DET"} \\ 0 & \text{sinon.} \end{cases}$$

$$f_{502}(i, y_i, y_{i-1}, x) = \begin{cases} 1 & \text{si } x_{i-1} = \text{"le" ET } y_{i-1} = \text{"DET"} \\ & \quad \text{ET } y_i = \text{"NC"} \\ 0 & \text{sinon.} \end{cases}$$

Modèle CRF linéaire

Formule (Lafferty et al. 2001)

Formule

$$P(y|x) = \frac{1}{Z(x)}.exp(\sum_i \sum_k \lambda_k f_k(i, y_i, y_{i-1}, x))$$

avec $Z(x)$ un facteur de normalisation

Logarithme

$$-logP(y|x) = logZ(x) - \sum_i \sum_k \lambda_k f_k(i, y_i, y_{i-1}, x)$$

Modèle CRF linéaire

Application à un exemple

k	f_k	λ_k
1	x_i="je" ET y_i="CLS"	-10
2	x_i="je" ET y_i="CC"	10
3	x_i="cours" ET y_i="NC"	-2
4	x_i="cours" ET y_i="V"	-1
5	y_{i-1}="CLS" ET y_i="V"	-10
6	y_{i-1}="CLS" ET y_i="NC"	5

x = je cours/y = CLS V

$$\begin{aligned} -log(P(y|x)) - logZ(x) &= (-10+0+0+0+0+0) \\ &+(0+0+0-1-10+0) \\ &= -21 \end{aligned}$$

Modèle CRF linéaire

Les propriétés des séquences de tokens

Propriétés de tokens

- Chaque token a un ensemble de propriétés (soit binaires ou textuelles)
- Ces propriétés sont choisies par l'utilisateur et calculées lors d'un prétraitement
- Les traits peuvent combiner les différentes propriétés des tokens

Exemples

- Propriétés binaires: le token commence par une majuscule, contient un chiffre, etc.
- Propriétés textuelles: la forme du token, suffixe de taille 3, etc.

Modèle CRF linéaire

Exemple de séquence des tokens prétraitée

position	forme	est token fusionné	contient un chiffre	suffixe de taille 2
10	,	-	-	-
11	soit	-	-	it
12	une	-	-	ne
13	augmentation	-	-	on
14	de	-	-	de
15	1_,_2	+	+	_2
16	%	-	-	-
17	par_rapport_au	+	-	rt
18	mois	-	-	is
19	précédent	-	-	nt

Jeu de traits pour l'étiquetage morphosyntaxique (inspiré de Tsuruoka et al., 2009)

Notations

- w_0 est le token courant
- w_k est le token à k positions du token courant
- t_0 et t_{-1} sont respectivement l'étiquette courante et l'étiquette précédente

Jeu de traits pour l'étiquetage morphosyntaxique (inspiré de Tsuruoka et al., 2009)

$w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}$	t_0
$w_{-1}w_0, w_0w_{+1}, w_{-1}w_{+1}$	t_0
suffixes de taille 1 à 5	t_0
préfixes de taille 1 à 5	t_0
w_0 contient un chiffre	t_0
w_0 contient un tiret	t_0
$w_0$commence par une majuscule	t_0
$w_0$est un token fusionné	t_0
forme en minuscule de w_0	t_0
t_{-1}	t_0

Couplage avec un lexique externe (Denis et Sagot, 2009)

Propriété provenant d'un lexique externe

- la classe d'ambiguité AC d'un token
- i.e. la concaténation des catégories grammaticales trouvées dans le lexique pour le token
- exemple : *par_rapport_au* (P+D), *précédent* (A_NC), *soit* (ADV_CS_VS)

Nouveaux traits pour CRF

Pour chaque type de lexique :

- trait interne: AC_0 avec t_0
- traits externes (1): AC_{-2},AC_{-1},AC_{+1},AC_{+2} avec t_0
- traits externes (2): AC_{-1}/AC_0, AC_0/AC_{+1}, AC_{-1}/AC_{+1} avec t_0

Les unités polylexicales dans FTB

Transformation du corpus

- Décomposition des tokens fusionnés en séquences de tokens simples (FTB+)
- Exemple: *par_rapport_au* (1 token) $\rightarrow$ *par rapport au* (3 tokens)
- 623 582 tokens dans FTB+ vs. 569 080 tokens dans FTB

Statistiques sur les unités polylexicales

- 87 479 tokens simples inclus dans une unité polylexicale ou un nombre (14% des tokens de FTB+)
- 2,7 tokens simples en moyenne par tokens fusionnés

Les unités polylexicales dans FTB

Types

- mots composés classiques (ex. *dans l' immédiat*, *acquis sociaux*, *en dehors de*)
- expressions verbales (ex. *fait l' objet*, *fait face*)
- fonctions politiques (*chancelier de l' Echiquier*)
- quelques entités nommées: noms d'organisation (*Société suisse de microélectronique et d' horlogerie*), noms de famille (*Strauss-Kahn*), noms de lieu (*Afrique du Sud*)

Grandes catégories manquantes

- Dates
- Noms de personne

Etiquetage avec identification des unités polylexicales

Principe

- Combiner segmentation et étiquetage
- On modifie le jeu d'étiquettes: à chaque token simple de FTB+, appartenant à un token étiqueté X de FTB, on associe soit l'étiquette $X + B$ (si début de token de FTB) soit l'étiquette $X + B$ (sinon)
- Cela revient à une tâche d'annotation classique (un token $\rightarrow$ une étiquette).

Etiquetage avec identification des unités polylexicales

Nouveau corpus

```
augmentation NC+B
de P+B
1 DET+B
, DET+I
2 DET+I
% NC+B
par P+D+B
rapport P+D+I
au P+D+I
mois NC+B
précédent ADJ+B
```

Nouvelles propriétés provenant de lexiques externes

Nouvelles propriétés d'un token

Pour chaque lexique d'unités polylexicales:

- `CAT`: catégorie grammaticale du mot composé auquel il appartient
- `STRUCT`: structure interne du mot composé auquel il appartient
- `SEM`: trait sémantique du mot composé auquel il appartient
- `POS`: position relative du token dans le mot composé

Prétraitement

- Utilisation des programmes d'Unitex (Paumier, 2003)
- Application glissante de grammaires locales de gauche à droite

Exemples de prétraitement avec ressources lexicales externes

FORME	CAT	STRUCT	POS	SEM
un	-	-	-1	-
gain	-	-	-1	-
de	-	-	-1	-
pouvoir	NC	NPN	0	-
d’	NC	NPN	1	-
achat	NC	NPN	2	-
de	-	-	-1	-
celles	-	-	-1	-
de	-	-	-1	-
la	-	-	-1	-
Banque	NPP	-	0	ORG
de	NPP	-	1	ORG
Chine	NPP	-	2	ORG

Nouveaux traits provenant de lexiques externes

CAT_0/POS_0	$\&t_0$
$STRUCT_0/POS_0$	$\&t_0$
SEM_0	$\&t_0$

Intégration d'une analyse lexicale préalable

Motivation

- Permettre un pré-filtrage des catégories grammaticales possibles par token
- Réaliser une pré-segmentation en unités multi-mots

Intégration à l'étiqueteur

- Analyse ambigüe de la séquence de tokens au moyen de ressources lexicales
- Représentation de l'analyse ambigüe par transducteurs (TFST)
- Pondération du TFST au moyen du modèle CRF appris
- Annotation par composition de transducteurs pondérés puis sélection du plus court chemin

Notations

Modèles CRF

- STD: traits classiques (indépendants de la langue)
- LEX: STD + traits utilisant les classes d'ambiguités des lexiques externes
- MWE: LEX + traits utilisant les informations sur les mots composés

LGTagger

- Sur FTB, analyse lexicale + CRF-LEX
- Sur FTB-MWE, analyse lexicale + CRF-MWE

Comparaison d'étiquetage sur FTB

NOM	REFERENCE	MODELE	DEV	TEST
TnT	(Brants 2000)	HMM	96,34	96,26
TreeTagger	(Schmid 1994)	Arbres de décision	96,45	96,35
SVMTool	(Giménez et al. 2004)	SVM	97,46	97,19
CRF-STD		CRF	(97,59)	97,39
MElt	(Denis et al. 2009)	MaxEnt	97,77	97,56
CRF-LEX		CRF	(97,95)	97,68
LGTagger		CRF	(98,00)	97,70

Comparaison d'étiquetage sur FTB+

NOM	MODELE	DEV	TEST
TreeTagger	Arbres de décision	89,71 (92,84)	
SVMTool	SVM	91,96 (94,20)	92,10 (94,67)
CRF-STD	CRF	(93,45) [95,44]	93,67 (95,83)
CRF-LEX	CRF	(93,69) [95,44]	93,91 (95,90)
CRF-MWE	CRF	(94,20) [95,95]	94,42 (96,42)
LGTagger	CRF	(94,17) [95,93]	94,34 96,34

Analyse d'erreurs

FTB (Sagot et al. 2009)

- erreurs classiques (ex. adjectif vs. participe passé)
- entités nommées
- nombres
- erreurs d'annotation manuelle du corpus

FTB+

- mots composés absents des ressources lexicales (entitées nommées)
- filtrage important de certaines catégories de noms composés (de type NA par exemple)

Conclusion

Consolidation de connaissances

- CRF très bon modèle statistique pour l'étiquetage morphosyntaxique
- Apprentissage très lent
- Le couplage avec des ressources lexicales riches améliore nettement les performances

Contributions

- Baisse sensible des performances si on effectue la segmentation en parallèle
- L'utilisation de ressources d'unités polylexicales améliore nettement les performances de segmentation
- L'apport de l'analyse préalable est mitigé

Perspectives

Ressources

- Ajuster mieux les ressources au corpus
- Ajouter un extracteur statistique d'entités nommées
- Annoter manuellement les unités polylexicales manquantes

Semi-CRF

- Tester le modèle semi-CRF plus adapté à la segmentation
- Combiner semi-CRF et transducteurs pondérés
- Super-chunking...

MERCI

Questions et remarques?