

Traitement Automatique de la Langue

Sources:

- <http://www.dptinfo.ens-cachan.fr/Conferences/gardent.pdf>
- <https://www.lattice.cnrs.fr/sites/itellier/cours-HMM-CRF-P3.pdf>

Plan

- Introduction
- Bref historique du Traitement Automatique des Langues Naturelles
- Etapes de l'analyse linguistique
- Applications

Introduction

Langages

- Un langage est un système incluant un ensemble de symboles, une syntaxe (pour former des expressions complexes à partir des symboles) et une sémantique (définissant le sens des expressions du langage)
- Exemples : lambda calcul, logique des prédictats, langages de programmation
- Les langages formels sont généralement peu ambigus : la syntaxe et le sens de chaque expression tendent à être uniques

Langues naturelles

Parlée (et écrite) par des humains

- Anglais, français, allemand, chinois, etc.

Deux différences importantes entre langages formels et langues naturelles :

- Multidimensionalité : le décodage d'une expression fait intervenir l'analyse syntaxique et sémantique mais aussi l'analyse phonétique, phonologique, morphologique, pragmatique (interaction avec le contexte) ainsi que le raisonnement basé sur les connaissances
- Combinatoire forte
 - ▶ Ambiguité : plusieurs analyses syntaxiques et/ou sémantiques possibles.
 - ▶ Paraphrases : le même contenu peut être exprimé de différentes façons.
 - ▶ Objectifs : gérer/réduire la combinatoire ; résoudre les ambiguïtés (analyse) ; faire les choix appropriés (génération)

Les niveaux d'analyse linguistique

Tous les niveaux d'analyse linguistique sont pertinents :

- Phonétique, Phonologie : sons / phonèmes / morphèmes
- Morphologie : morphèmes / mots
- Syntaxe : mots / constituants
- Sémantique : syntaxe / sens littéral
- Pragmatique : sens littéral, contexte / sens en contexte

Ambiguité

L'ambiguité est présente à tous les niveaux linguistiques.

- Phonologique: Le même *signal sonore* peut avoir plusieurs interprétations possibles :
Recognise speech ou Wreck a nice peach ??
- Sémantique lexicale: Le même *mot* peut dénoter différents objets.
étoile : célébrité ou astre?
- Partie du discours (catégorie morphosyntaxique): Le même *mot* peut avoir différentes catégories.
la : pronom, nom ou déterminant ?
- Syntaxe: La même *phrase* peut avoir plusieurs analyses syntaxiques.
Jean regarde (la fille avec un télescope)
Jean ((regarde la fille) avec un télescope)
- Sémantique phrastique: La même *phrase* peut avoir plusieurs analyses sémantiques.
La belle ferme la porte

Les applications du TAL

Traitent du texte et utilisent des connaissances linguistiques.

- Les interfaces vocales
- La reconnaissance de l'écriture manuelle (handwriting recognition)
- La correction orthographique
- La recherche d'information (K. Spark-Jones, 1972) e.g., les moteurs de recherche (Google, Yahoo, etc.)
- La traduction automatique e.g., Google Translate
- Les systèmes de dialogue homme-machine
- L'enseignement des langues assisté par ordinateur
- La détection d'opinions (à partir des blogs, des pages web, des réseaux sociaux)
- etc.

Bref historique du TAL

Les années 50

- Traduction automatique
- Russe/Anglais
- George Town University, Washington system:
 - ▶ Traduit un texte court en 1954

1964: le rapport ALPAC

- Évalue les recherches en traduction automatique
- Conclut que la traduction automatique est impossible dans un futur proche
- Arrêt des financements
- Début des recherches fondamentales en TAL

La traduction mot-à-mot donne de très mauvais résultats

Des *Connaissances linguistiques* sont nécessaires

Les années 60 : TAL, Linguistique et Logique

- 1957 : *Syntactic Structures* de Noam Chomsky
 - ▶ Définition formelle des grammaires et des langues
 - ▶ Base pour l'analyse syntaxique des langues
- 1967 : sémantique procédurale de Woods
 - ▶ Une approche procédurale du sens
 - ▶ Base pour le traitement automatique du sens
- PTQ de Montague
 - ▶ sémantique formelle pour la langue naturelle.
 - ▶ Base pour un traitement logique du sens

Les premiers succès

- 1970 – TAUM Meteo
Traduction automatique de bulletins météos (Canada)
- 1970s – SYSTRAN: TA; maintenant devenu Babelfish
- 1973 – Lunar
Interrogation d'un système expert sur les analyses d'échantillons minéraux lunaires
- 1973 – SHRDLU (T. Winograd)

Les années 80: les approches symboliques

- Premières grosses grammaires informatiques
- Linguistique + logique
- Manque de robustesse
- Peu d'applications

Les années 80s : Corpus et Ressources

- Espace disque peu cher
- Texte numérisé disponible
- Evaluation des systèmes sur des données réelles (passage à échelle)
- 1994 – Le [British National Corpus](#) (BNC) est publié
Un corpus balancé de l'anglais britannique
- 1990s – [WordNet](#) (Fellbaum & Miller)
Un thesaurus informatique développé par des psycholinguistes
- 2000s – Utilisation du World Wide [Web comme corpus](#)

Historique – Résumé

- Années 50: Traduction automatique – débuts du TAL
- 1964 Rapport ALPAC
- Années 60: Linguistique formelle (Chomsky, Montague) comme base pour le TAL. Applications basées sur des techniques linguistiques (Eliza, shrdlu) – Chomsky (grammaires formelles, analyseurs syntaxiques); sémantique procédural (Woods) . Approches limitées à des domaines restreint. Non portables.
- Années 70: Premières applications
- Années 80: Approches symboliques. Applications utilisent des connaissances linguistiques et encyclopédiques extensives. Manquent de robustesse.
- Années 90 et plus: Premiers corpus, approches statistiques, apprentissage automatique. Applications utilisent corpus de grande taille et méthodes statistiques

Etapes de l'analyse linguistique:

Analyse lexicale

Difficultés de la Segmentation / Normalisation

- ▶ Les écritures sans segmentation (chinois, thaï...)
- ▶ S'accommoder des ambiguïtés typographiques :
 - ▶ . : dans *etc.*, dans *20.3*, dans *enst.com*, dans ..., dans *TF.1...*
 - ▶ ' : dans *jusqu'à*, dans *aujourd'hui*, dans *3'4*, dans *Sotheby's* ou *Floc'h* ...
 - ▶ - : dans *Jean-Michel*, dans *donne-t-il*, dans *06-04-62-26-16-23*, dans *1914-1918*, dans *-1.2 %...*
 - ▶ sans parler de l'espace lui-même
- ▶ Déetecter et normaliser les variantes typographiques : *France-Inter* *France-inter* et *France Inter* ; *États-Unis* et *Etats-unis* et *Etats-Unis...*
- ▶ "Reconnaître" les chiffres, dates, durées, nombres, montants, numéros (de téléphone, de carte bleue), les scores...
- ▶ "Faire avec" les mots inconnus, les emprunts, les coquilles...

Le niveau lexical

- ▶ **But** : identifier les éléments lexicaux, leur structure et leurs caractéristiques ; regrouper les formes d'une même famille.
- ▶ **Moyen** : accès lexical direct, analyse morphologique (i.e. décomposition en *morphèmes*, à partir desquels les propriétés d'une forme sont calculées).
- ▶ **Outils** : un lexique, une description des morphèmes et des procédures de décomposition/recomposition associées.
- ▶ **Difficultés** : taille du lexique, vitesse d'accès et d'analyse, représentation du lexique, traitement des mots composés.
- ▶ **Résultat** : une représentation linéaire ou arborescente du mot, ses caractéristiques morpho-syntactiques, une représentation de sa signification, un représentant de sa famille.

Le traitement lexical : résultat

- ▶ *le* - det. masc. sing., /lə/ ; pron. pers. masc. sing., /lə/
- ▶ *président* - vrb 3pers. plur. prés. ind./ subjonctif [présid+ent],
<présider(X), présider(X,Y)>, /pʁezid+ət/ ; nom masc. sing.,
← présider : action de X, <president(X)>, /pʁezidā/
- ▶ *des* - det. masc./fem. plur., /dε+z/ ; prep. contr. *de les* ...
- ▶ *antialcooliques* - adj. masc./fem. plur. [anti+alcool+ique+s], ←
alcoolique : s'opposer à X, antialcoolique(X), /ɑ̃tialkɔlikə+z/ ; nom.
masc. sing. [anti+alcool+ique+s], ← antialcoolique (adj) : être X,
antialcoolique(X), /ɑ̃tialkɔlikə+z/
- ▶ *mangeait* - vrb (1,3) pers. sing. imp. ind., [mang+e+ait],
<manger(X),manger(X,Y)>, /māʒε+t/
- ▶ *pomme* - nom fem. sing., [pomme], <pomme(X),fruit(X),golden(X)...>,
/pɔmə/
- ▶ ...

Etiquetage morpho-syntaxique

Catégories syntaxiques

- Les mots peuvent être regroupés en classes d'après leur comportement syntaxique.
- 8 grandes catégories : nom, verbe, pronom, préposition, adverbe, conjonction, adjetif et article.
- Autre catégories utilisées: eg Penn Treebank (45 étiquettes), Susanne (353 étiquettes).

Le jeu d'étiquettes du Penn Treebank (1)

CC	Coord Conjuncn	<i>and, but, or</i>	NN	Nom, sing. or mass	<i>dog</i>
CD	Cardinal number	<i>one, two</i>	NNS	Nom, plural	<i>dogs</i>
DT	Article	<i>the, some</i>	NNP	Proper nom, sing.	<i>Edinburgh</i>
EX	Existential there	<i>there</i>	NNPS	Proper nom, plural	<i>Orkneys</i>
FW	Foreign Word	<i>mon dieu</i>	PDT	Prearticle	<i>all, both</i>
IN	Préposition	<i>of, in, by</i>	POS	Possessive ending	<i>'s</i>
JJ	Adjectif	<i>big</i>	PP	Personal pronom	<i>I, you, she</i>
JJR	Adj., comparative	<i>bigger</i>	PP\$	Possessive pronom	<i>my, one's</i>
JJS	Adj., superlative	<i>biggest</i>	RB	Adverbe	<i>quickly</i>
LS	List item marker	<i>1, One</i>	RBR	Adverbe, comparative	<i>faster</i>
MD	Modal	<i>can, should</i>	RBS	Adverbe, superlative	<i>fastest</i>

Le jeu d'étiquettes du Penn Treebank (2)

RP	Particle	<i>up, off</i>	WP\$	Possessive-Wh	<i>whose</i>
SYM	Symbol	<i>+, %, &</i>	WRB	Wh-adverbe	<i>how, where</i>
TO	"to"	<i>to</i>	\$	Dollar sign	<i>\$</i>
UH	Interjection	<i>oh, oops</i>	#	Pound sign	<i>#</i>
VB	verbe, base form	<i>eat</i>	"	Left quote	<i>', "</i>
VBD	verbe, past tense	<i>ate</i>	"	Right quote	<i>', "</i>
VBG	verbe, gerund	<i>eating</i>	(Left paren	<i>(</i>
VBN	verbe, past part	<i>eaten</i>)	Right paren	<i>)</i>
VBP	Verbe, non-3sg, pres	<i>eat</i>	,	Comma	<i>,</i>
VBZ	Verbe, 3sg, pres	<i>eats</i>	.	Sent-final punct	<i>. ! ?</i>
WDT	Wh-article	<i>which, that</i>	:	Mid-sent punct.	<i>: ; — ...</i>
WP	Wh-pronom	<i>what, who</i>			

Étiquettage : Définition

- L'étiquettage associe un catégorie syntaxique unique à chaque mot d'un texte
- Exemple:
“ The/DT guys/NNS that/WDT make/VBP traditional/JJ hardware/NN
are/VBP really/RB being/VBG obsoleted/VBN by/IN
microprocessor-based/JJ machines/NNS ./, said/VBD Mr./NNP
Benton/NNP ./.”

Étiquettagé et ambiguïté

- L'étiquettagé vise à lever l'ambiguité morpho-syntaxique
 - ▶ The *back/JJ* door
 - ▶ On my *back/NN*
 - ▶ With the voters *back/RB*
 - ▶ He promised to *back/VB* the bill
- Brown Corpus

1 cat.	2 cat.	3 cat.	4 cat.	5 cat.	6 cat.	7 cat.
35340	3760	264	61	12	2	1

NLTK: Python Natural Language ToolKit

- Projet libre source
<http://nltk.sourceforge.net>
- Développeurs : Steven Bird, Ewan Klein, Ed Loper
- NLTK est un ensemble de modules python qui implémentent des algorithmes pour le TAL :
 - ▶ traitement des expressions régulières
 - ▶ extraction de phrases et de mots
 - ▶ étiquettage
 - ▶ analyse locale
 - ▶ analyse syntaxique
 - ▶ etc.

Méthodes d'étiquettage dans NLTK

- Règles spécifiées manuellement
 - ▶ Étiquettage basé sur les expressions régulières (la forme d'un mot est utilisée pour deviner sa catégorie syntaxique)
- Approche statistique (Modèles de Markov cachés)
 - ▶ Étiquettage à partir d'unigrammes : associe à chaque mot, l'étiquette la plus fréquente pour ce mot
 - ▶ Étiquettage à partir de bigrammes : associe à chaque mot, l'étiquette la plus probable étant donnée l'étiquette la plus fréquente pour ce mot et l'étiquette du mot précédent
- Étiquettage à la Brill : apprentissage de règles d'étiquettage par transformations

Evaluer un étiquetteur

Phrase	Référence	Étiqueteur
The	at	at
President	nn-tl	nn-tl
said	vbd	vbd
he	pps	pps
will	md	md
ask	vb	vb
Congress	np	np
to	to	to
increase	vb	*nn*
grants	nns	nns
to	in	*to*
states	nns	nns

L'étiqueteur a correctement étiquetté 10 mots sur 12.
Précision = 10/12 ou 83%

Evaluer un étiquetteur

- Idée de base : comparer la sortie de l'étiquetteur avec une sortie produite manuellement, la *référence*
- les meilleurs étiquetteurs automatiques ont une précision de 96-97% pour un jeu d'étiquettes relativement petit i.e., celui du Penn treebank
- l'accord entre humains est de seulement 97%
- Un bon étiquetteur par unigrammes a une précision d'environ 90-91%

Approche probabiliste

- Aussi appelé étiquettage par les Modèles de Markov cachés
- On cherche à identifier parmi toutes les séquences c_1^n de catégories possibles pour la phrase d'entrée m_1^n , la séquence dont la probabilité est la plus grande

$$c_1^n = \operatorname{argmax}_{c_1^n} P(c_1^n | m_1^n)$$

- On utilise la règle de Bayes et la propriété de Markov pour transformer cette équation en une équation pour laquelle les probabilités peuvent être estimées à partir d'un corpus annoté.

Utilisation de la règle de Bayes

$$P(x|y) = \frac{P(y|x).P(x)}{P(y)}$$

$$P(c_1^n|m_1^n) = \frac{P(m_1^n|c_1^n)P(c_1^n)}{P(m_1^n)}$$

$$c_1^n = argmax_{c_1^n} \frac{P(m_1^n|c_1^n)P(c_1^n)}{P(m_1^n)}$$

$$c_1^n = argmax_{c_1^n} P(m_1^n|c_1^n)P(c_1^n)$$

Utilisation de la propriété de Markov

On simplifie encore le modèle :

$$P(m_1^n | c_1^n) = \prod_{i=1}^n P(m_i | c_i)$$

$$P(c_1^n) = \prod_{i=1}^n P(c_i | c_{i-1})$$

$$c_1^n = \operatorname{argmax}_{c_1^n} P(c_1^n | m_1^n) = \operatorname{argmax}_{c_1^n} \prod_{i=1}^n P(m_i | c_i) P(c_i | c_{i-1})$$

Probabilités de transition

- Probabilités de transition de catégories $P(c_i|c_{i-1})$
 - ▶ Un déterminant précède souvent un nom
 - ▶ La probabilité $P(NN|DT)$ est élevée
- On estime ces probabilités en comptant les cas pertinents dans un corpus annoté

$$P(c_i|c_{i-1}) = \frac{C(c_{i-1}c_i)}{C(c_{i-1})}$$

Probabilités d'émission

- VBZ (verbe à la 3ème personne du singulier du présent) peut être “est”
- On estime cette probabilité en comptant les cas pertinents dans un corpus annoté

$$P(m_i|c_i) = \frac{C(c_i, m_i)}{C(c_i)}$$

$$P(is|VBZ) = \frac{C(VBZ, is)}{C(VBZ)}$$

Exemple

Secretariat/NNP is/VBZ expected/VBZ to/TO race/VB tomorrow/NN

People/NNS continue/VBP to/TO inquire/VB the/DT reason/NN for/IN
the/DT race/NN for/IN outer/JJ space/NN

- “race” est un verbe dans la première phrase mais un nom dans la seconde.
- Probabilité pour “race” d’être un verbe/nom dans le premier exemple:

$$P(\text{race} \text{ is } VB) = P(VB|TO)P(\text{race}|VB)$$

$$P(\text{race} \text{ is } NN) = P(NN|TO)P(\text{race}|NN)$$

Exemple

$$P(NN|TO) = 0.021$$
$$P(VB|TO) = 0.34$$

$$P(\text{race}|NN) = 0.00041$$
$$P(\text{race}|VB) = 0.00003$$

$$\begin{aligned} P(\text{race is } VB) &= P(VB|TO)P(\text{race}|VB) \\ &= 0.34 \times 0.00003 = 0.00001 \\ P(\text{race is } NN) &= P(NN|TO)P(\text{race}|NN) \\ &= 0.021 \times 0.00041 = 0.000007 \end{aligned}$$

Limites des étiquetteurs par n-grammes

- Taille de la table de n-grammes (*modèle de langage*): centaines de millions d'entrées
- Contexte limité aux étiquettes; une information sur le mot précédent ou suivant est souvent utile

Étiquettagé à la Brill

- Un système de règles...
- ... où les règles sont acquises automatiquement à partir d'un corpus

Étiquettag par transformation

Idée de base :

- ① on étiquette chaque mot avec son étiquette la plus probable (étiquettag par unigramme)
- ② on définit un ensemble fini de règles transformationnelles (*Dans le contexte C, remplace C1 par C2*)
- ③ On applique chacune de ces règles au corpus
- ④ on sélectionne la meilleure règle i.e., celle qui améliore le plus l'étiquettag
- ⑤ On ré-étiquette les données en utilisant la nouvelle règle

On répète 1-3 jusqu'à ce que les améliorations soient très faibles.

Sortie: ensemble ordonné de transformations; procédure d'étiquettag.

Les transformations de Brill

Les transformations utilisées sont toutes de la forme

Change étiquette A en étiquette B si:

- le mot précédent/suivant est étiquetté z
- le second mot précédent/suivant est étiquetté z
- un des trois mots précédents/suivants est étiquetté z
- le mot précédent est étiquetté z et le mot suivant est étiquetté w
- le mot précédent/suivant est étiquetté z et le second mot précédent/suivant est étiquetté w

Exemple de transformation

Remplace NN par VB si la catégorie précédente est TO

... is/VBZ expected/VBN to/TO race/NN tomorrow/NN

⇒

... is/VBZ expected/VBN to/TO race/VB tomorrow/NN

Selectionner la meilleure transformation

- la meilleure transformation est la transformation avec le bénéfice net le plus élevé
- le bénéfice net d'une transformation est le nombre d'étiquettes incorrectes corrigées moins le nombre d'étiquettes correctes modifiées

Etiquettes morpho-syntaxiques universelles

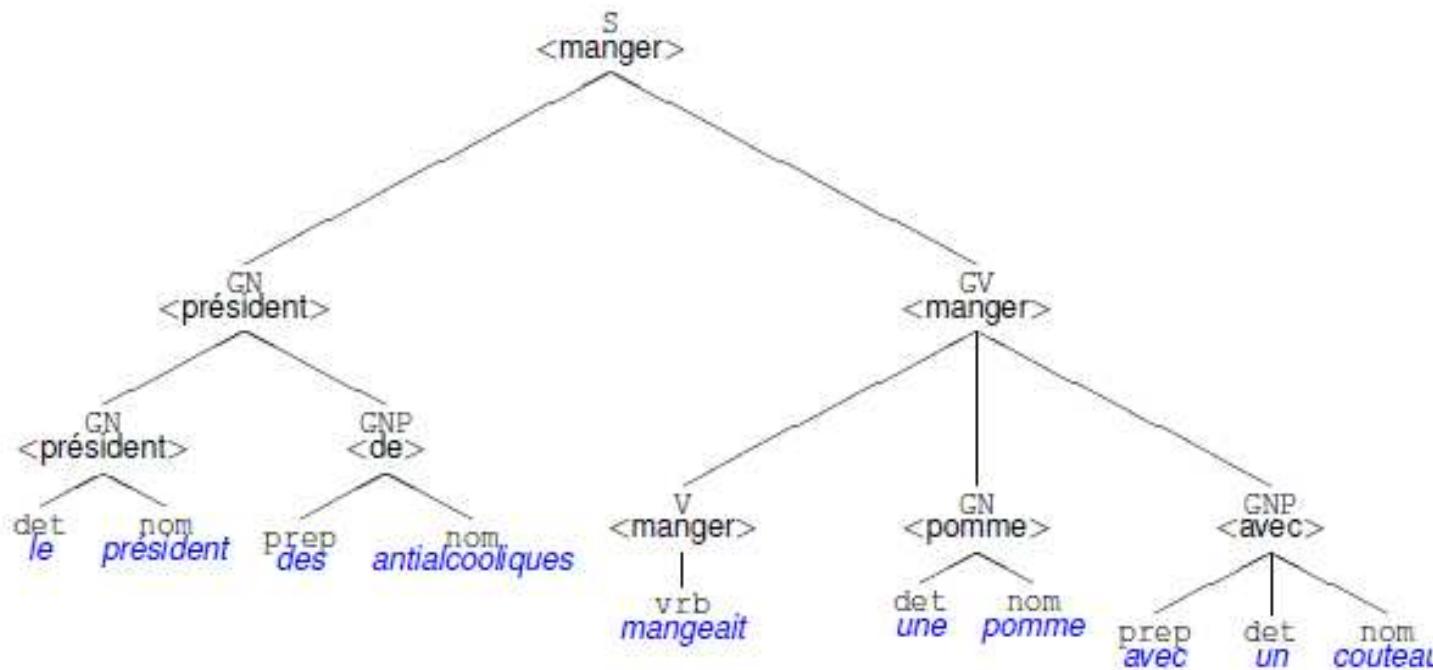
Etiquette Penn TreeBank	Description	Etiquette Universelle
CC	conjunction, coordinating	CONJ
CD	cardinal number	NUM
DT	determiner	DET
EX	existential there	DT
FW	foreign word	X
IN	conjunction, subordinating or preposition	ADP
JJ	adjective	ADJ
JJR	adjective, comparative	ADJ
JJS	adjective, superlative	ADJ
LS	list item marker	X
MD	verb, modal auxillary	VERB
NN	noun, singular or mass	NOUN
NNS	noun, plural	NOUN
NNP	noun, proper singular	NOUN
NNPS	noun, proper plural	NOUN
PDT	predeterminer	DET
POS	possessive ending	PRT
PRP	pronoun, personal	PRON
PRPDOL	pronoun, possessive	PRON
RB	adverb	ADV
RBR	adverb, comparative	ADV
RBS	adverb, superlative	ADV
RP	adverb, particle	PRT
SYM	symbol	X
TO	infinitival to	PRT
UH	interjection	X
VB	verb, base form	VERB
VBD	verb, 3rd person singular present	VERB
VBG	verb, non-3rd person singular present	VERB
VBN	verb, past tense	VERB
VBP	verb, past participle	VERB
VBZ	verb, gerund or present participle	VERB
WDT	wh-determiner	DET
WP	wh-pronoun, personal	PRON
WPDOL	wh-pronoun, possessive	PRON
WRB	wh-adverb	ADV
,	punctuation mark, sentence closer	.
,	punctuation mark, comma	.
:	punctuation mark, colon	.
(contextual separator, left paren	.
)	contextual separator, right paren	.

Analyse syntaxique

Le niveau syntaxique

- ▶ **But** : identifier les composants syntaxiques (syntagmes), leur fonction, et les relations qu'ils entretiennent entre eux.
- ▶ **Moyen** : analyse syntaxique, qui fournit une représentation arborescente des composants de l'énoncé.
- ▶ **Outils** : un analyseur syntaxique, c'est-à-dire un formalisme de description des règles syntaxiques, des règles valides pour un (sous)-langage donné, et un système d'analyse (un parseur) capable d'exploiter ces règles.
- ▶ **Difficultés** : compromis entre richesse de description, vitesse d'analyse, et prolifération des ambiguïtés, complexité des phénomènes à décrire, robustesse aux entrées "bruitées" (coquilles, casse...).
- ▶ **Résultat** : un (ou des) arbres syntaxiques représentant la phrase.

Le traitement syntaxique : résultat



L'ambiguïté lexicale

Un des principaux problèmes de l'analyse syntaxique est l'ambiguïté.

Ambiguïté lexicale :

- ▶ *souris* : formes verbales de *sourir*, nom féminin singulier et pluriel ;
- ▶ *petit* : adjectif ou nom masculin singulier ;
- ▶ *la* : déterminant ou pronom personnel féminin singulier, nom masculin ;
- ▶ *mousse* : formes verbales de *mousser*, nom masculin, nom féminin ;

Plus la description lexicale est précise, plus l'ambiguïté est grande : *monter* (*monter un escalier*, *monter un cheval*, *monter une pièce*, ...).

Cette ambiguïté n'est pas seulement statique, mais également *dynamique* : les phénomènes syntaxiques de *translation* rendent ambigus adjectifs et participes passés (emploi nominal) : *ces affreux se sont enfuis*

L'ambiguïté syntaxique

- ▶ *La petite brise la glace* ;
- ▶ *La troupe monte Molière* vs *Le jockey monte Belino* ;
- ▶ *Elle mange une pomme avec les doigts* vs *Elle mange une pomme avec la peau* ;
- ▶ *Elle mange une glace à la fraise* vs *Elle mange une glace à la plage* ;
- ▶ *C'est la fille du cousin qui boit* ;
- ▶ *Il a parlé de déjeuner avec Paul* ;

La désambiguïsation est possible au niveau sémantique ou pragmatique ; chaque raffinement de la grammaire accroît l'ambiguïté.

Context Free Grammars

A context free grammar $G = (N, \Sigma, R, S)$ where:

- ▶ N is a set of non-terminal symbols
- ▶ Σ is a set of terminal symbols
- ▶ R is a set of rules of the form $X \rightarrow Y_1 Y_2 \cdots Y_n$
for $n \geq 0$, $X \in N$, $Y_i \in (N \cup \Sigma)$
- ▶ $S \in N$ is a special start symbol

Context Free Grammars: Example

$N = \{S, NP, VP, Adj, Det, Vb, Noun\}$

$\Sigma = \{fruit, flies, like, a, banana, tomato, angry\}$

$S = 'S'$

$R =$

$S \rightarrow NP\ VP$

$NP \rightarrow Adj\ Noun$

$NP \rightarrow Det\ Noun$

$VP \rightarrow Vb\ NP$

$Adj \rightarrow fruit$

$Noun \rightarrow flies$

$Vb \rightarrow like$

$Det \rightarrow a$

$Noun \rightarrow banana$

$Noun \rightarrow tomato$

$Adj \rightarrow angry$

Context Free Grammars: Left-most derivations

Left-most derivation is a sequence of strings s_1, \dots, s_n where

- ▶ $s_1 = S$ the start symbol
- ▶ $s_n \in \Sigma^*$, meaning s_n is only terminal symbols
- ▶ Each s_i for $i = 2 \dots n$ is derived from s_{i-1} by picking the left-most non-terminal X in s_{i-1} and replacing it by some β where $X \rightarrow \beta$ is a rule in R .

For example: [S],[NP VP],[Adj Noun VP], [fruit Noun VP], [fruit flies VP],[fruit flies Vb NP],[fruit flies like NP], [fruit flies like Det Noun], [fruit flies like a], [fruit flies like a banana]

Context Free Grammars: Syntactic Tree

$S \rightarrow NP\ VP$

$NP \rightarrow Adj\ Noun$

$NP \rightarrow Det\ Noun$

$VP \rightarrow Vb\ NP$

-

$Adj \rightarrow fruit$

$Noun \rightarrow flies$

$Vb \rightarrow like$

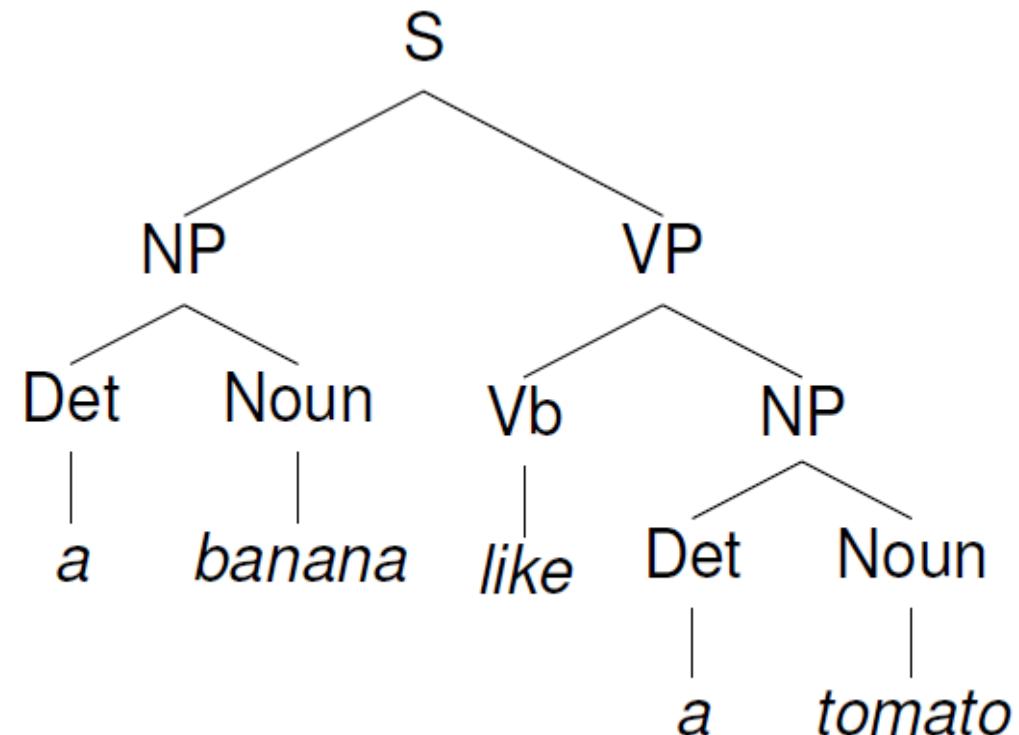
$Det \rightarrow a$

$Noun \rightarrow banana$

$Noun \rightarrow tomato$

$Adj \rightarrow angry$

...



Context Free Grammars: Syntactic Tree

$S \rightarrow NP\ VP$

$NP \rightarrow Adj\ Noun$

$NP \rightarrow Det\ Noun$

$VP \rightarrow Vb\ NP$

-

$Adj \rightarrow fruit$

$Noun \rightarrow flies$

$Vb \rightarrow like$

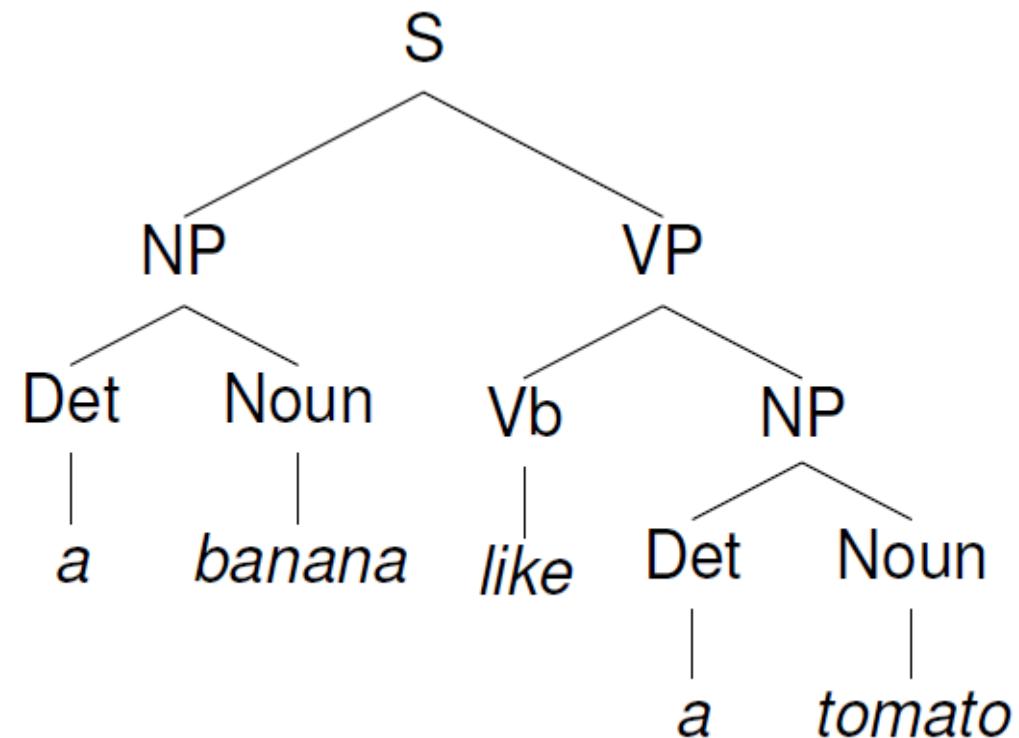
$Det \rightarrow a$

$Noun \rightarrow banana$

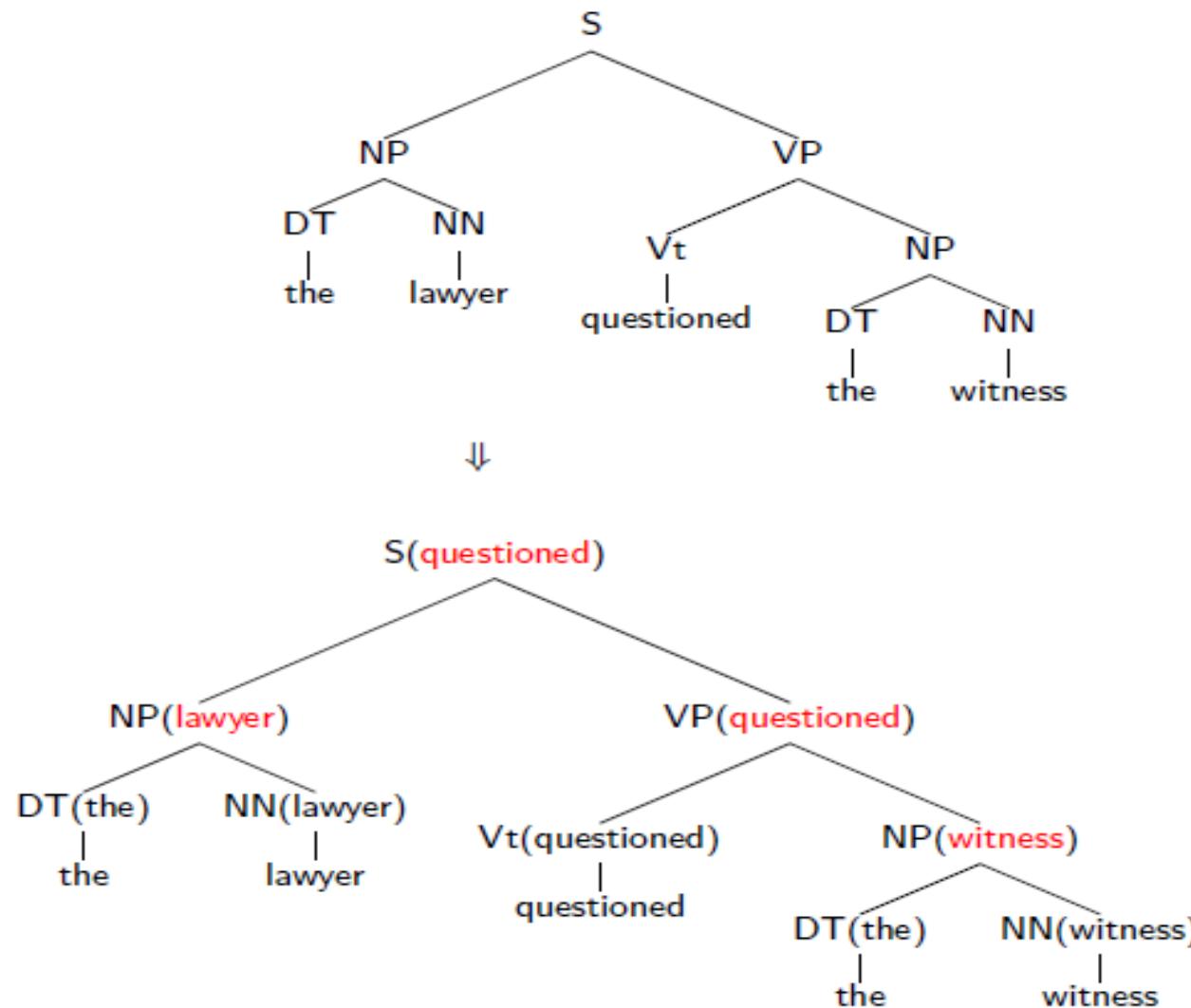
$Noun \rightarrow tomato$

$Adj \rightarrow angry$

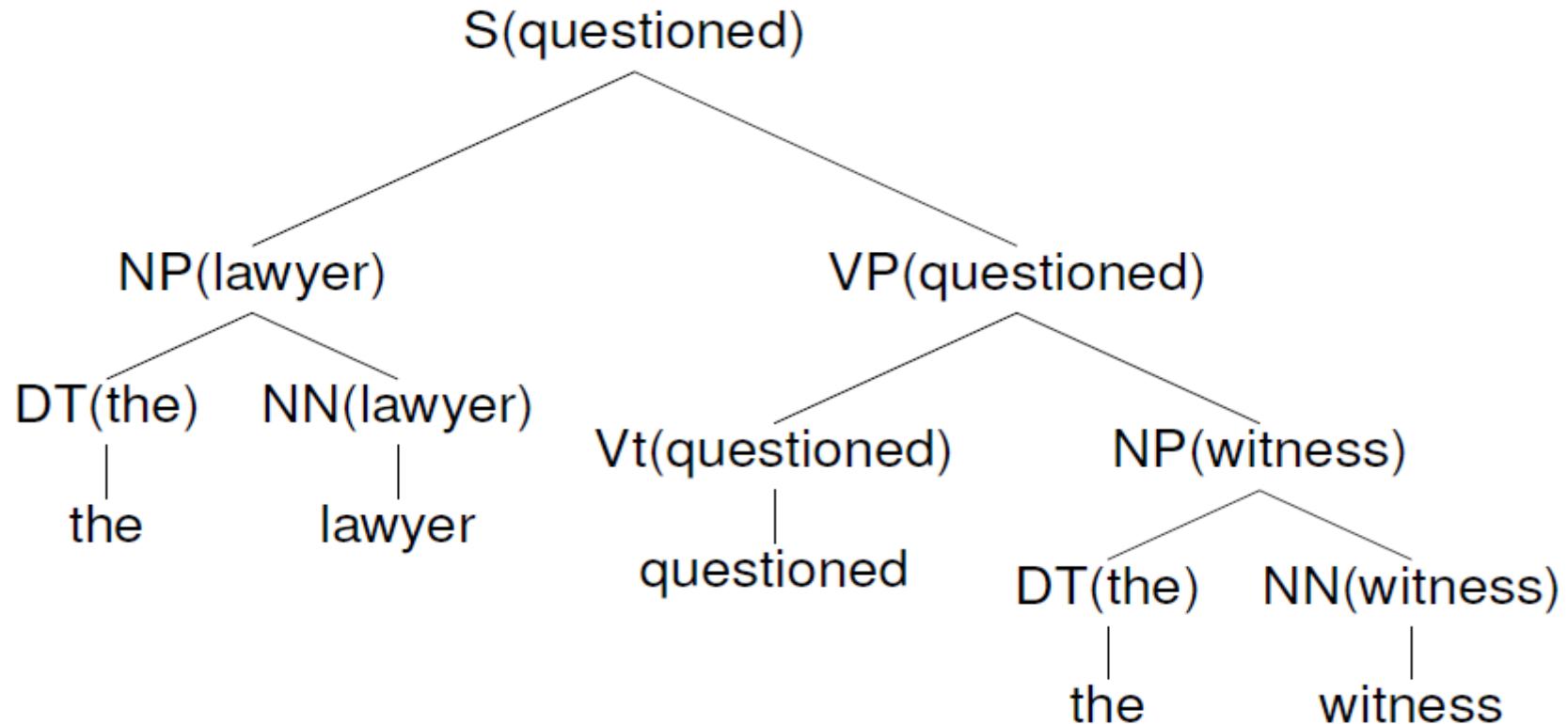
...



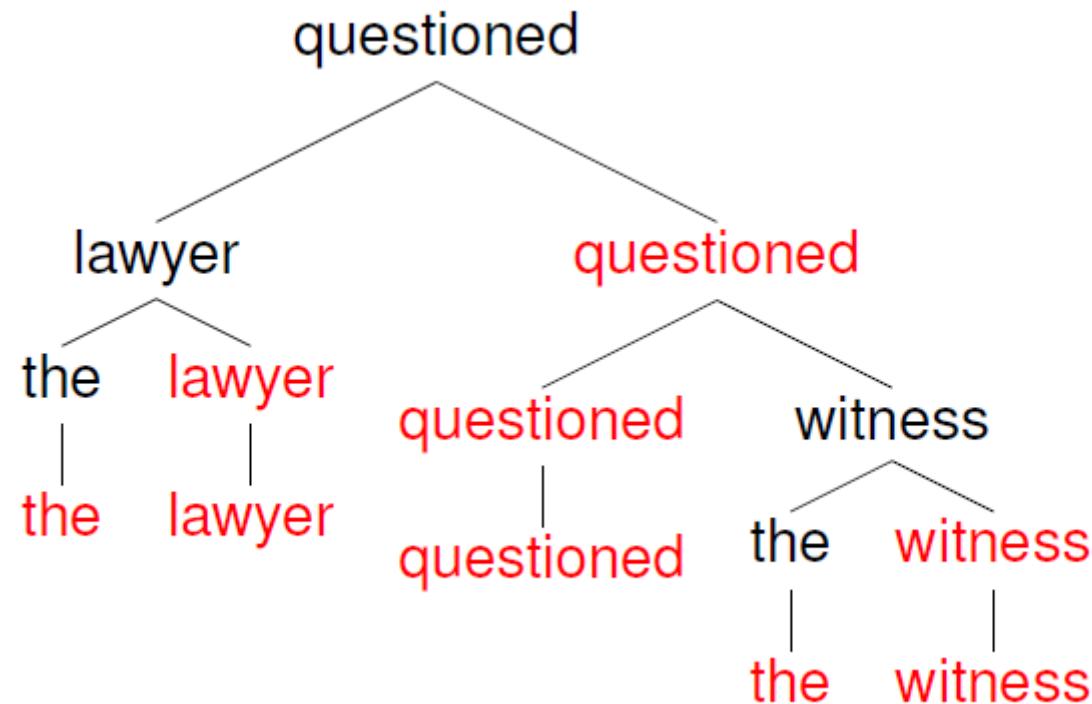
Dependency Parsing: Capturing the relation between words in a sentence



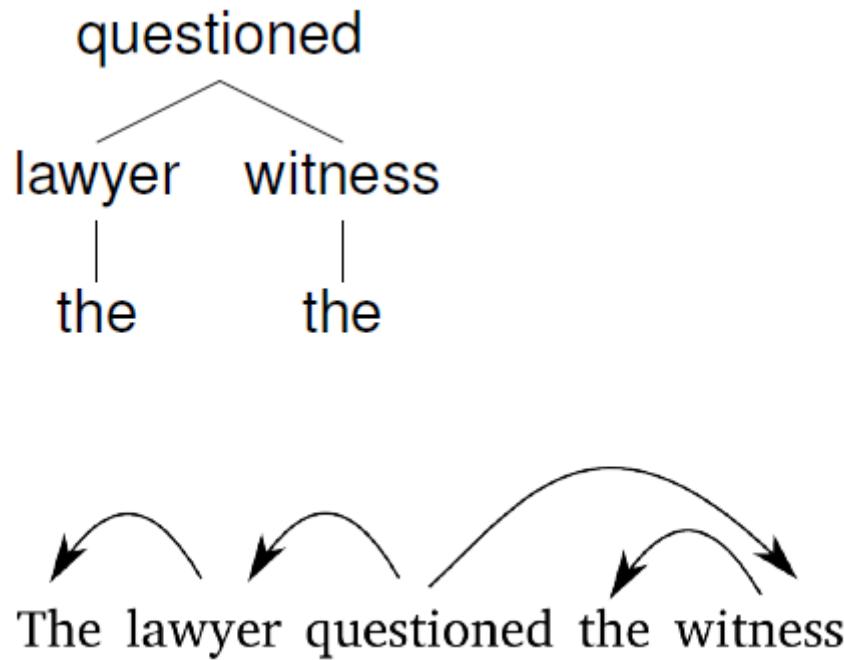
Dependency Parsing: Capturing the relation between words in a sentence



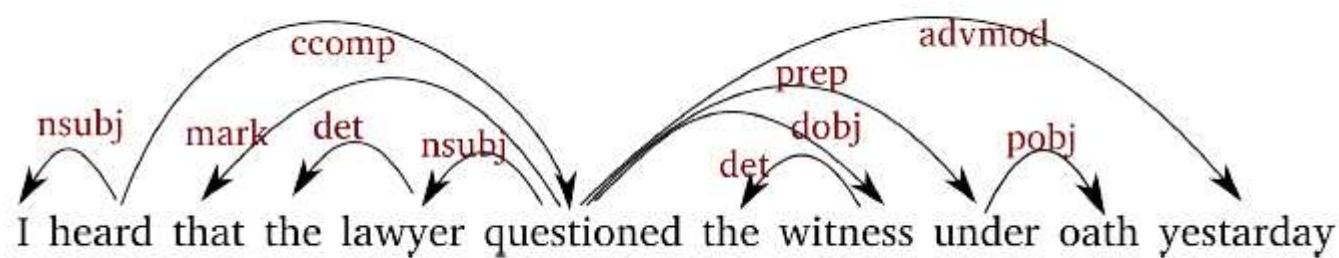
Dependency Representation



Dependency Representation



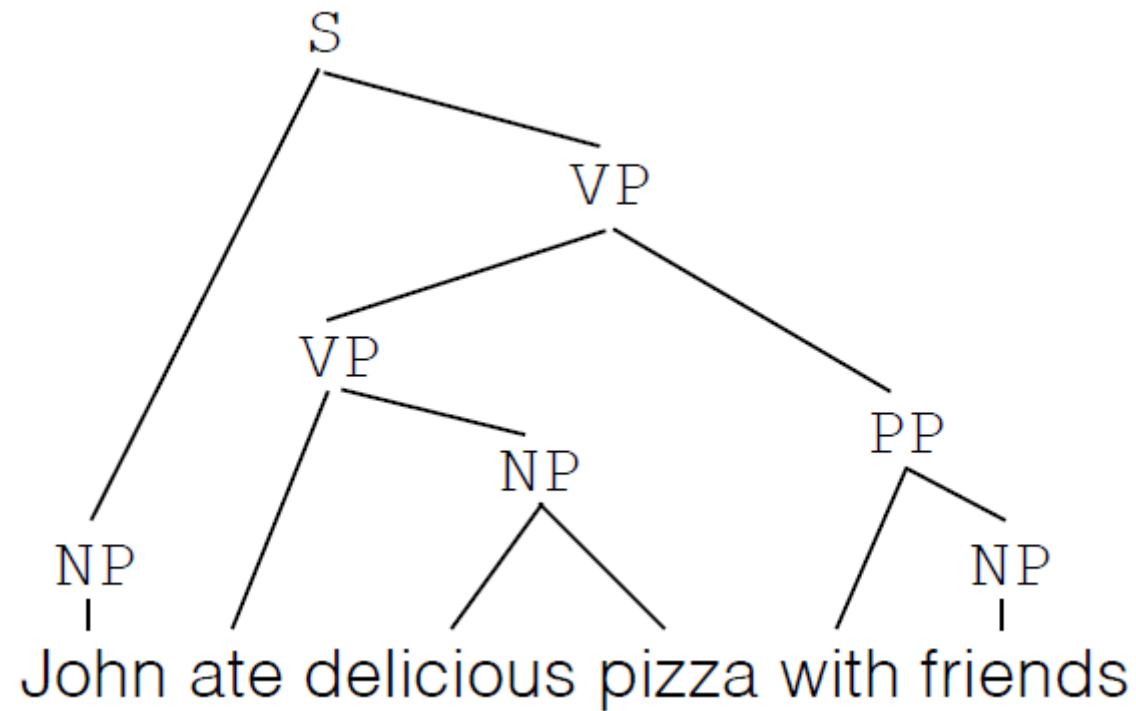
Dependency Representation



Dependency relations (Stanford Dependencies):

- **nsubj**: nominal subject
- **mark**: subordinating conjunction
- **ccomp**: clausal complement
- **det**: determiner
- **prep**: preposition prepositional modifier
- **dobj**: direct object
- **advmod**: adverbial modifier
- **pobj**: object of a preposition

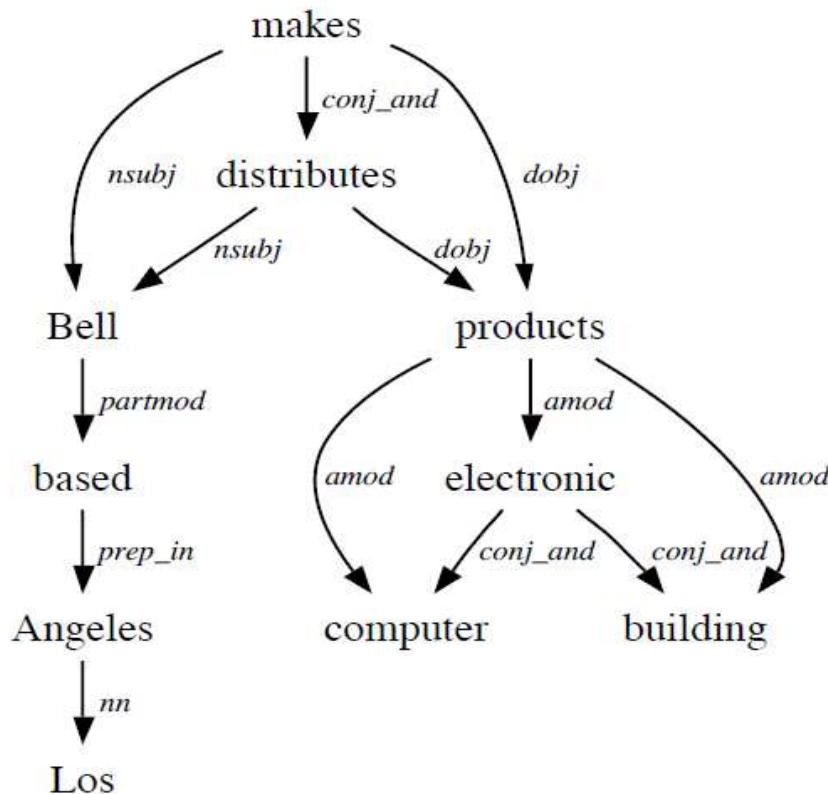
Analyse syntaxique: Arbre syntaxique



Analyse syntaxique: Graphe de dépendance

Phrase:

Bell, based in Los Angeles, makes and distributes electronic, computer and building products.



- nsubj** (makes-8, Bell-1)
- nsubj** (distributes-10, Bell-1)
- vmod** (Bell-1, based-3)
- nn** (Angeles-6, Los-5)
- prep** in (based-3, Angeles-6)
- root** (ROOT-0, makes-8)
- conj** and (makes-8, distributes-10)
- amod** (products-16, electronic-11)
- conj** and (electronic-11, computer-13)
- amod** (products-16, computer-13)
- conj** and (electronic-11, building-15)
- amod** (products-16, building-15)
- dobj** (makes-8, products-16)
- dobj** (distributes-10, products-16)

Analyse syntaxique à base de règles: Exemple

Phrase: *Quelle journée*

Règle: *pour reconnaître la relation entre un déterminant interrogatif et un nom (la relation entre le déterminant interrogatif "quelle" et le nom commun "journée")*

```
@DetInt :: (@AdjPren) {0-1} (@Substantif | $L_DIVERS-L_DIVERS_DATE) :  
SYNTACTIC_RELATION:  
+SecondUngovernedBy(trigger.1, right.2, "ANY")  
+GenderAgreement(trigger.1, right.2)  
+NumberAgreement(trigger.1, right.2)  
+CreateRelationBetween(trigger.1, right.2, "DetIntSub")
```

La première ligne correspond à la règle respectivement :

- un déclencheur @DetInt
- un contexte droit (vide)
- et un contexte gauche (@AdjPren) {0-1} (@Substantif | \$L_DIVERS-L_DIVERS_DATE) séparés les uns des autres par un ::

Les deux lignes suivantes représentent des contraintes d'accord. La dernière ligne est la directive de création de la relation nommée DetIntSub entre trigger.1 (la première partie du déclencheur : @DetInt) et right.2 (la deuxième partie du contexte droit : (@Substantif | \$L_DIVERS-L_DIVERS_DATE))

Etiquettes universelles pour les relations de dépendance syntaxique

Etiquette Universelle	Description
root	the head of a sentence
nsubj	nominal subject
nsubjpass	passive nominal subject
csubj	clausal subject
csubjpass	clausal passive subject
dobj	direct object
iobj	indirect object
ccomp	clausal complement
xcomp	open clausal complement
nmod	nominal modifier
advmod	adverbial modifier
advcl	adverbial clause modifier
neg	negation
appos	apposition
amod	adjectival modifier
acl	clausal modifier of a noun (adjectival clause)
det	determiner
case	case marking
vocative	addressee
aux	auxiliary verb
auxpass	passive auxiliary
cop	copula verb
mark	subordinating conjunction
expl	expletive
conj	conjunct
cc	coordinating conjunction
discourse	discourse element
compound	relation for marking compound words
name	names
mwe	multiword expressions that are not names
foreign	text in a foreign language
goeswith	two parts of a word that are separated in text
list	used for chains of comparable elements
dislocated	dislocated elements
parataxis	parataxis
remnant	remnant in ellipsis
reparandum	overridden disfluency
punct	punctuation
dep	unspecified dependency

Extraction d'information

- L'information d'aujourd'hui sur support informatique est :
 - massive
 - complexe et hétérogène
 - soumise à des contraintes de temps réel
- **Extraction d'Information (IE)** : conversion du texte en données structurées répondant à des questions factuelles

QUI A FAIT QUOI A QUI QUAND OU COMMENT

- Applications : recherche, indexation, aide à la décision, veille, question/réponse, construction de ressources ...

Extraction d'information

Exemple:

Un raid aérien a fait au moins 11 morts et 12 blessés sur le village de Menakro le mardi 12 février

ENTITES	SEGMENTS	REPRESENTANTS
DATE	le mardi 12 février	12/02/03
LIEU	le village de Menakro	Menakro
FAIT	Un raid aérien	Attaque militaire
IMPACT	au moins 11 morts	Pertes humaines (<50)
	(au moins) 12 blessés	Dommages humains (<50)

Extraction d'Entités Nommées

- La tâche d'Extraction d'Information a mis en évidence l'intérêt de reconnaître les Entités Nommées :

- **Qu'est-ce que c'est une Entité Nommée ?**

...tous les éléments du langage qui font référence à une entité unique et concrète, appartenant à un domaine spécifique (humain, économique, géographique, etc.)

...noms propres mais aussi expressions de temps et de quantité

(MUC-7, Chinchor 1998)

Extraction d'Entités Nommées

De manière générale, il s'agit de noms propres pouvant être classés

- dans des catégories prédéfinies
 - ENAMEX : organisation, lieu, personne
 - TIMEX : dates, expressions temporelles
 - NUMEX : valeurs monétaires, pourcentage, ...
- dans des catégories spécifiques à un domaine
 - biologie : espèces, protéines, gènes, etc.
 - médecine : médicaments, conditions médicales, etc.
 - mais aussi noms de bateau, modèles d'avion, etc.

Extraction d'Entités Nommées

Exemple

Né à Paris[lieu] le 21 octobre 1944[date], Jean-Pierre Sauvage[personne] a effectué sa thèse à l'Université de Strasbourg[lieu][organisation,lieu] sous la direction de Jean-Marie Lehn[personne]. Après un post-doctorat à Oxford[organisation,lieu], il revient en France[lieu] et effectue sa carrière au CNRS[organisation] qu'il intègre en 1971[date] et devient directeur de recherche au CNRS[organisation] en 1979[date]. Jean-Pierre Sauvage[personne] travaille à l'Institut de science et d'ingénierie supramoléculaire[organisation] (CNRS[organisation]/Université de Strasbourg[lieu][organisation,lieu]). Il a également reçu la médaille de bronze en 1978[date] et celle d'argent du CNRS[organisation] en 1988[date].

On peut reconnaître

- les entités nommées, imbriquées ou non
- les types associés aux entités, parfois ambiguës

Extraction d'Entités Nommées

Obstacle important en TAL :

- Majorité des mots inconnus d'un corpus
- Porteurs d'informations importantes
- Similaires aux groupes nominaux complexes avec beaucoup de variation
 - (Wikipedia EN) Carl XVI Gustaf of Sweden, Carl XVI Gustaf, Carl Gustaf Folke Hubertus, King Carl Gustaf, His Majesty Carl XVI Gustaf, King of Sweden, Carl Gustaf
 - (Wikipedia FR) Barack Obama, Barack Hussein Obama II, Barack Obama Jr., Obama, président Obama, président Barack Obama
- Acronymes peuvent être similaires aux mots : *OTAN, Laser, Radar*
- Nécessitent plusieurs analyses

Approches pour l'Extraction d'Entités Nommées

- Utilisation de dictionnaires ou de listes existantes
- Définition d'automates
- Analyse statistique ou reconnaissance par apprentissage automatique

Difficultés dans l'Extraction d'Entités Nommées

- La portée des classes : Clint Eastwood, l'épouse Chirac, les frères Cohen, les démocrates, les Boeings, Bison futé
- La coordination : Barack et Michelle Obama, M. et Mme Obama
- L'imbrication : Université de Strasbourg
- Les frontières : l'équipe de Nantes, le Palais Bourbon, monsieur Hollande/le président Hollande, le couple Obama
- Les variantes : l'équipe de Nantes/le stade nantais/les canaris/les nantais/Nantes/FCN
- La polysémie : Clint Eastwood (acteur, réalisateur, producteur, mais aussi chanteur jamaïquain, chanson, personne de film), Leclerc (maréchal, homme d'affaire, Char, supermarché)

Approches symboliques pour l'Extraction d'Entités Nommées

- Projection de dictionnaires
 - On retrouve les entités nommées connues
 - Catégorisation des entités nommées

- Utilisation des majuscules

Alan Turing, Metro Goldwin Mayer, Nobody Can Beat the Wiz

- Indice insufisant : le premier mot des phrases est généralement en majuscule...
- Problème de la limite à droite
Institut national de recherche en informatique et en automatique
- *Organisation des Nations Unies efficace*
- Solution : utilisation de grammaires des EN et du lexique

Approches symboliques pour l'Extraction d'Entités Nommées

- Projection de dictionnaires
 - On retrouve les entités nommées connues
 - Catégorisation des entités nommées
- Utilisation des majuscules

Alan Turing, Metro Goldwin Mayer, Nobody Can Beat the Wiz

- Indice insufisant : le premier mot des phrases est généralement en majuscule...
- Problème de la limite à droite
Institut national de recherche en informatique et en automatique
- *Organisation des Nations Unies efficace*
- Solution : utilisation de grammaires des EN et du lexique

Approches symboliques pour l'Extraction d'Entités Nommées

Projection de dictionnaires

Utile pour reconnaître des catégories d'entités nommées précises
Mais

- inutiles si trop petits
- sources d'ambiguïté si trop grands

et de toute façon, ils ne sont pas exhaustifs !

En général :

- Utilisation de dictionnaires d'EN combinés à des indices externes ou internes identifiés manuellement ou automatiquement
- Pour les lieux : utilisation de dictionnaires

Approches symboliques pour l'Extraction d'Entités Nommées

Projection de dictionnaires

Utile pour reconnaître des catégories d'entités nommées précises
Mais

- inutiles si trop petits
- sources d'ambiguïté si trop grands

et de toute façon, ils ne sont pas exhaustifs !

En général :

- Utilisation de dictionnaires d'EN combinés à des indices externes ou internes identifiés manuellement ou automatiquement
- Pour les lieux : utilisation de dictionnaires

Approches symboliques pour l'Extraction d'Entités Nommées

Approches à base de règles

- hors-contexte : utilisation d'indices internes
- contextuelle : utilisation d'indices externes

Approches symboliques pour l'Extraction d'Entités Nommées

Approche hors-contexte

- Utilisation des caractéristiques de la séquence

Les entités ont une structure interne :

- **Luc** Besson, **F.** Hollande, **H.** Clinton
- **docteur** Jean Dupond, **maître** Durant, **président** Obama
- Sherwood **Forest**, Hollywood **Boulevard**, Place de l'étoile, **aéroport** d'Orly
- **groupe** Vivendi, **société** Général, Airbus **group**

- Utilisation d'indices internes à l'entité

- Majuscule, prénoms, abréviation de prénoms
- Mots classifiant des métiers des lieux, des organisations
- ...

Approches symboliques pour l'Extraction d'Entités Nommées

Approche contextuelle

- Hypothèse : existence d'un contexte facilitant l'identification d'entités nommées et leur catégorisation
- Utilisation du contexte locaux des entités :
 - Personne : titre, métier, grade, ...
juge van Ruymbeke, **docteur** Freud, **monsieur** Chirac, **général** De Gaulle
 - Organisation : statut, activité, ..
la **filiale** de PSA, la **compagnie** Ryanair, le **motoriste** Safran, **constructeur aéronautique** Airbus
 - Lieux :
la **ville** de Rennes, le **fleuve** amazone, la **comète** Tchouri, le **sud** de Paris, **basé à** Lyon, **lac** Baïkal
 - Mais aussi contexte spécifique :
Transcription of the cotB, cotC, and cotX **genes**
la **sonde** Rosetta, le robot **Philae**

Identification des Entités Nommées

Problèmes

- conflit entre indices internes et externes
La société Yves Saint-Laurent, le groupe Hugo Boss, la société Hughes Aircraft
→ On priviléie l'indice externe
- Ambiguité du contexte :
 - *All American Bank vs. All State Police*
 - *JFK* (mais aussi *Charles De Gaulle*)
→ Un contexte plus large doit être utilisé
- Ambiguité de la coordination
 - *C&A, H&M, Pratt & Whitney vs. Apple et Samsung*

Approches symboliques pour l'Extraction d'Entités Nommées

Approche à base de règles: Exemples

Exemple 1: Règle pour la reconnaissance des noms de journaux français "Libération", "Le Monde Diplomatique", etc.

```
Libération:::ORGANIZATION:  
Monde:Le:Diplomatique::ORGANIZATION:  
Monde:Le:de l'Education::ORGANIZATION:  
Monde:Le::ORGANIZATION:  
Courrier::International::ORGANIZATION:  
Canard::Enchaîné::ORGANIZATION:
```

Exemple 2: Règle pour la reconnaissance des noms de personnes

```
@Firstname:[(@Title|@FunctionTitle)?]:  
((de|la|le)? T_A1){1-2}:PERSON:N_PERSON  
T_A1:[(@Title|@FunctionTitle)]:T_A1{0-2}:PERSON:N_PERSON  
T_A1:(T_A1|T_Amh){0-2}:, @FunctionTitle:PERSON:N_PERSON
```

Approches statistiques pour l'Extraction d'Entités Nommées

Utilisation de méthodes d'étiquetage séquentiel (par apprentissage)

- ① Données annotées selon le représentation BIO(/IOB)
- ② Apprentissage d'un modèle (HMM, CRF, etc.) sur les données annotées
- ③ Utilisation du modèle pour étiqueter les données selon la représentation BIO
- ④ Post-traitement pour interpréter la représentation BIO

Approches statistiques pour l'Extraction d'Entités Nommées

Représentation BIO

- Chaque mot est associé à une classe
 - **B** (Begin), **I** (Inside), **O** (Outside)
 - ou en prenant en compte la catégorie sémantique :
 - Personne : **B-PERS** (Begin), **I-PERS** (Inside)
 - Organisation : **B-ORG** (Begin), **I-ORG** (Inside)
 - ...
 - **O** (Outside)

Approches statistiques pour l'Extraction d'Entités Nommées

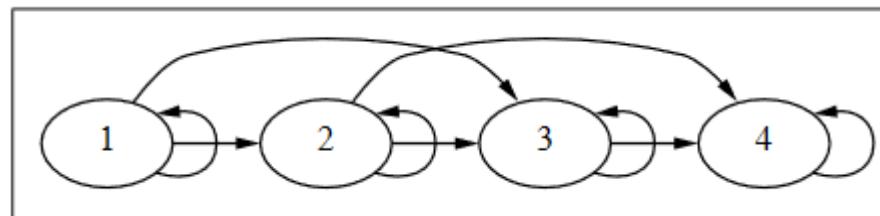
Exemple

Né		Né	O
à		à	O
Paris	LOC	Paris	B-LOC
le		le	O
21	DATE	21	B-DATE
octobre	DATE	octobre	I-DATE
1944	DATE	1944	I-DATE
,		,	O
Jean-Pierre	PERS	Jean-Pierre	B-PERS
Sauvage	PERS	Sauvage	I-PERS
a		a	O
effectué		effectué	O
sa		sa	O
thèse		thèse	O

Approches statistiques pour l'Extraction d'Entités Nommées

Modèle d'apprentissage avec des HMM

- Un HMM est un processus permettant d'engendrer une séquence
- Une séquence est une suite d'observations sur un HMM en fonctionnement
 - Exemple: HMM gauche-droite à quatre états



Algorithme : Génération d'une séquence par un HMM

début

$t \leftarrow 1$

 Choisir l'état initial $q_1 = s_i$ avec la probabilité π_i

tant que $t \leq T$ faire

 Choisir l'observation $o_t = v_k$ avec la probabilité $b_i(k)$

 Passer à l'état suivant $q_{t+1} = s_j$ avec la probabilité a_{ij}

$t \leftarrow t + 1$

fin tant que

fin

Approches statistiques pour l'Extraction d'Entités Nommées

Modèle d'apprentissage avec des HMM

- Les HMM sont des modèles probabilistes d'émission de séquences, discrètes ou continues, ils sont utilisés en classification bayésienne
- L'algorithme forward-backward permet de connaître la probabilité qu'un HMM ait émis une séquence
- L'algorithme de Viterbi permet de connaître la suite des états du HMM qui a la plus forte probabilité d'avoir émis une séquence
- L'algorithme de Baum-Welsh permet d'ajuster les paramètres d'un HMM au maximum de vraisemblance à partir d'un ensemble de séquences d'apprentissage

Approches statistiques pour l'Extraction d'Entités Nommées

Modèle d'apprentissage avec des CRF

- Définition d'un modèle probabiliste décrivant des caractéristiques de surface spécifiques aux entités nommées comme les CRF
- CRF - Conditional Random Field :
 - Objectif : maximiser $p(t|w)$ sans calculer de modèle $p(w)$ permet l'utilisation d'un ensemble de *features* plus important
 - Modèle graphique (CRF linéaire)

$$p(t|w) = \frac{\prod_{i=2}^N \exp(\sum_k \lambda_k f_k(t_{i-1}, t_i, w, i))}{\sum_{t'} \exp(\prod_{i=2}^N \exp(\sum_k \lambda_k f_k(t'_{i-1}, t'_i, w, i)))}$$

- Les *features* f_k doivent être définies par l'utilisateur
- Les paramètres du modèle (λ_k) sont estimés sur des données d'entraînement

t: Observations (Annotations), w: Variables

Les CRF

Premières propriétés des CRF

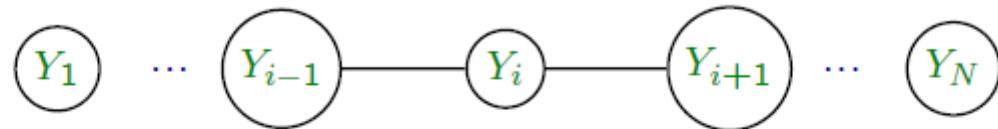
- modèle markovien : soit un graphe sur les Y , $p(Y_i|X)$ ne dépend que de X et des Y_j ($i \neq j$) avec lesquels Y_i est relié dans le graphe
- dans ce cas, on a (Hammersley-Clifford 71) :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in \mathcal{C}} \psi_c(y_c, x)$$

- \mathcal{C} est l'ensemble des cliques du graphe sur Y
- y_c : valeurs des variables de y sur la clique c
- $Z(x)$ un coefficient de normalisation
- chaque $\psi_c(y_c, x)$ est une fonction de potentiels
- les CRF sont des modèles graphiques markoviens non dirigés (donc non génératifs)
- il reste à définir un graphe sur les annotations Y_i

Les CRF sur les séquences

Le graphe des CRF linéaires (sur les chaînes)



La formule

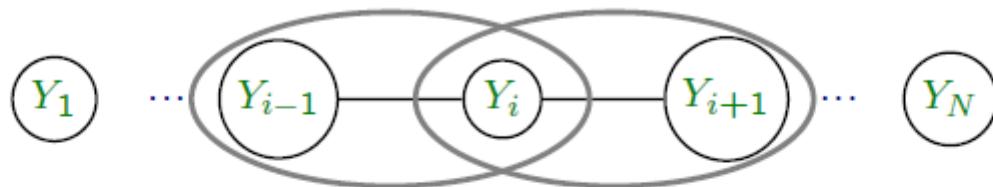
- proposition de (Lafferty, McCallum et Pereira 01) :

$$p(y|x) = \frac{1}{Z(x)} \prod_{i=2}^N \exp \left(\sum_k \lambda_k f_k(y_{i-1}, y_i, x, i) \right)$$

- chaque f_k est une fonction “feature” donnée par l’utilisateur
- c’est le même ensemble de f_k qui sert pour chaque clique
- chaque λ_k est un poids initialement inconnu (paramètres du modèle) associé à f_k

Les CRF sur les séquences

Les CRF "linéaires"



- implicite : les données X sont accessibles partout (on ne cherche pas à les générer comme dans un HMM)
- les cliques sont les couples (Y_{i-1}, Y_i) (en gris)
- exemples de features $f_k(y_{i-1}, y_i, x, i)$ à la position i :
 - * $f_k(y_{i-1}, y_i, x, i) = 1$ si $x_{i-1} \in \{la, une\}$ et $y_{i-1} = Det$ et $y_i = Nom$
= 0 sinon
 - * $f_{k'}(y_{i-1}, y_i, x, i) = 1$ si $\{M., Mme, Melle\} \cap \{x_{i-3}, \dots, x_{i-1}\} \neq \emptyset$
et $y_i = EN$
= 0 sinon

Approches statistiques pour l'Extraction d'Entités Nommées

Exemple d'utilisation de patrons de features

Né	VER:ppter	naître	O
à	PRP	à	O
Paris	NAM	Paris	B-LOC
le	DET:ART	le	O
21	NUM	@card@	B-DATE
octobre	NOM	octobre	I-DATE
1944	NUM	@card@	I-DATE
,	PUN	,	O
Jean-Pierre	NAM	Jean-Pierre	B-PERS
Sauvage	NAM	Sauvage	I-PERS
a	VER:pres	avoir	O
effectué	VER:ppter	effectuer	O
sa	DET:POS	son	O
thèse	NOM	thèse	O
à	PRP	à	O
l'	DET:ART	le	O
Université	NOM	université	B-ORG
de	PRP	de	I-PERS
Strasbourg	NAM	Strasbourg	I-ORG
sous	PRP	sous	O
la	DET:ART	le	O
direction	NOM	direction	O
de	PRP	de	O
Jean-Marie	NAM	Jean-Marie	B-PERS
Lehn	NAM	Lehn	I-PERS
.	SENT	.	O

Patron de features :
forme fléchie, étiquette morpho-syntaxique,
lemme, et étiquette EN du mot courant

Approches statistiques pour l'Extraction d'Entités Nommées

Exemple d'utilisation de patrons de features

Né	VER:ppter	naître	O
à	PRP	à	O
Paris	NAM	Paris	B-LOC
le	DET:ART	le	O
21	NUM	@card@	B-DATE
octobre	NOM	octobre	I-DATE
1944	NUM	@card@	I-DATE
,	PUN	,	O
Jean-Pierre	NAM	Jean-Pierre	B-PERS
Sauvage	NAM	Sauvage	I-PERS
a	VER:pres	avoir	O
effectué	VER:ppter	effectuer	O
sa	DET:POS	son	O
thèse	NOM	thèse	O
à	PRP	à	O
l'	DET:ART	le	O
Université	NOM	université	B-ORG
de	PRP	de	I-PERS
Strasbourg	NAM	Strasbourg	I-ORG
sous	PRP	sous	O
la	DET:ART	le	O
direction	NOM	direction	O
de	PRP	de	O
Jean-Marie	NAM	Jean-Marie	B-PERS
Lehn	NAM	Lehn	I-PERS
.	SENT	.	O

Patron de features :
étiquette morpho-syntaxique, lemme et
étiquette EN du mot courant
étiquette morpho-syntaxique et étiquette EN
du mot précédent

Approches statistiques pour l'Extraction d'Entités Nommées

Modèle d'apprentissage avec des CRF

Processus:

- Données annotées utilisés comme exemple $((w, y))$
- Définition des *features* ou des patrons de *features* ($f_k(\dots)$)
- Apprentissage des poids du CRF permettant d'obtenir un modèle (λ_k)
- Application du modèle sur de nouvelles données en cherchant la séquence d'annotations y qui maximise $p(t|w)$

Bilan:

- CRF : meilleures résultats pour les tâches correspondant à des annotations sur des séquences
- Autres possibilités : sans étiquetage séquentiel : arbres de décision, SVM, etc.

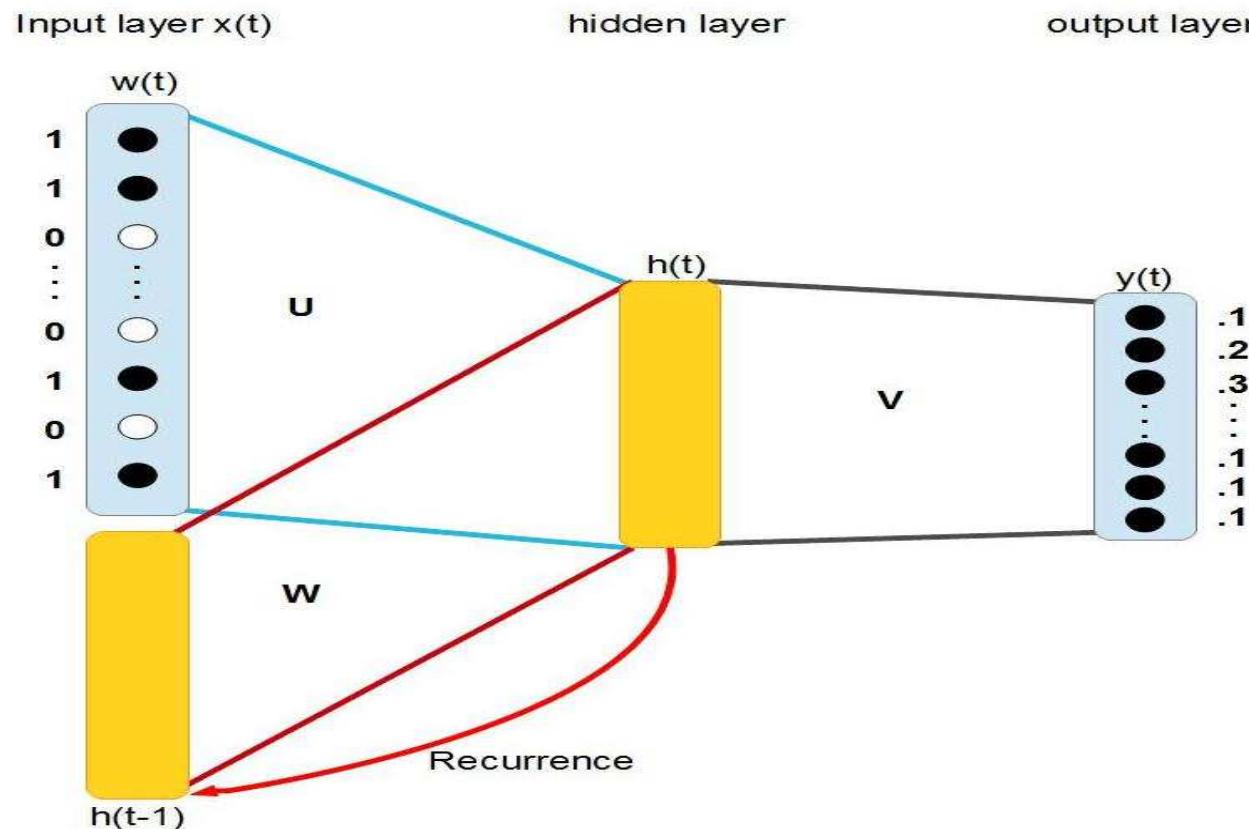
Performances des différentes approches pour l'Extraction d'Entités Nommées

- Utilisation de règles
 - Règles lisibles, évolution des systèmes par ajout de lexique, mais coût de la description
 - Surtout adaptés à la langue écrite
 - Rappel & précision > 90%
- Apprentissage de modèles
 - Modèles numériques, arbre de décision,... difficilement modifiables, mais coût de la description faible (nécessite un corpus d'apprentissage)
 - Surtout adaptés à la langue orale, mais aussi bonne performances sur les textes de spécialité
 - Rappel entre 50 et 90%
- Systèmes mixtes : avantages et inconvénients des deux

Mais performances variables suivant les entités nommées et le nombre de catégories

Approches neuronales pour l'Extraction d'Entités Nommées

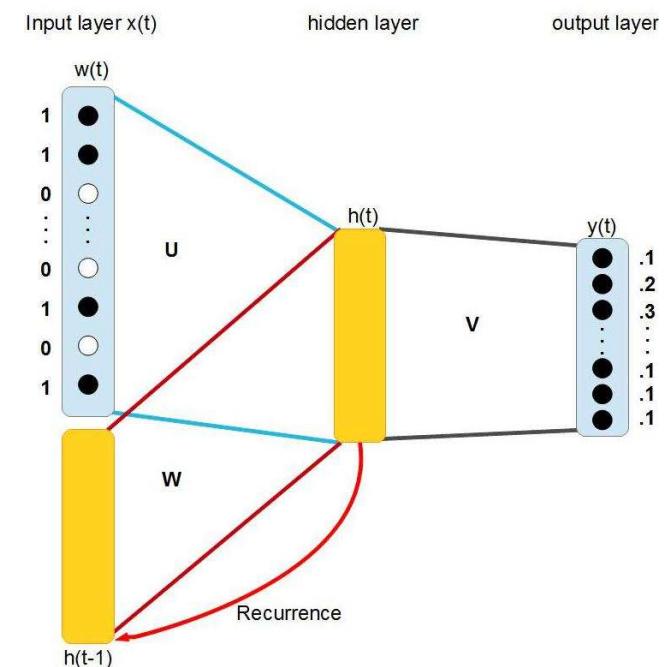
Architecture d'un modèle d'apprentissage à base de RNNs



Approches neuronales pour l'Extraction d'Entités Nommées

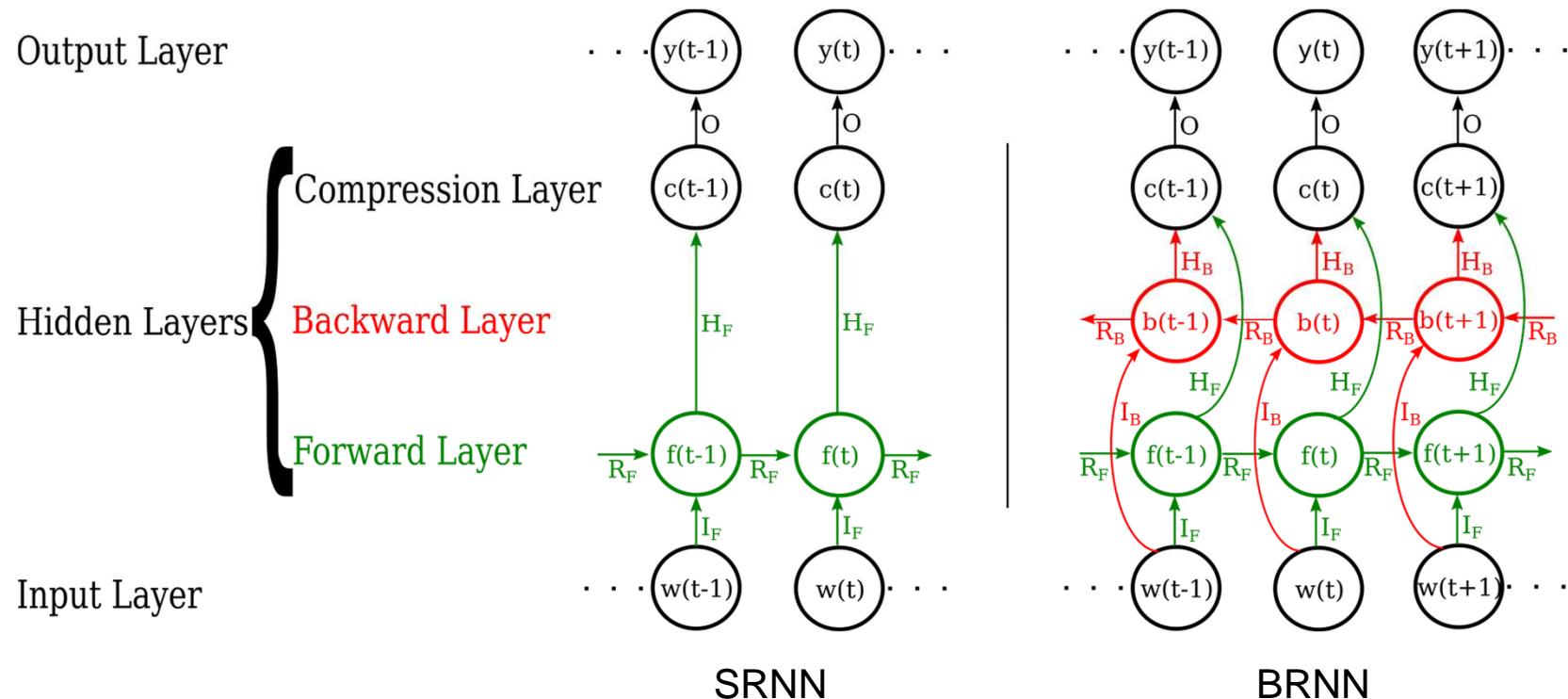
Algorithme d'apprentissage:

- The training is performed using Stochastic Gradient Descent (SGD)
- We go through all the training data iteratively, and update the weight matrices U , W and V online (after processing every word)
- Training is performed in several epochs



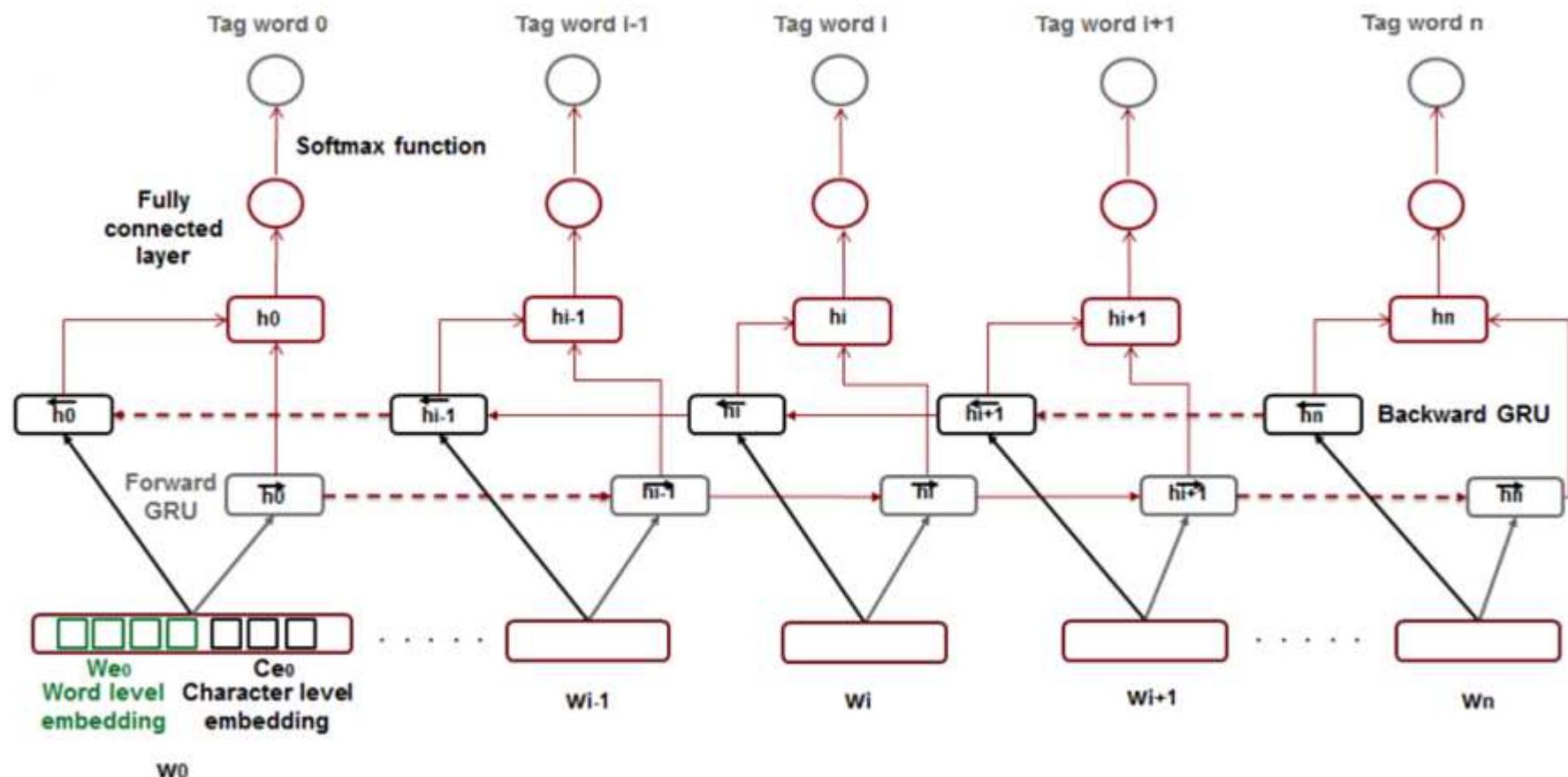
Approches neuronales pour l'Extraction d'Entités Nommées

RNNs bi-directionnels:



Approches neuronales pour l'Extraction d'Entités Nommées

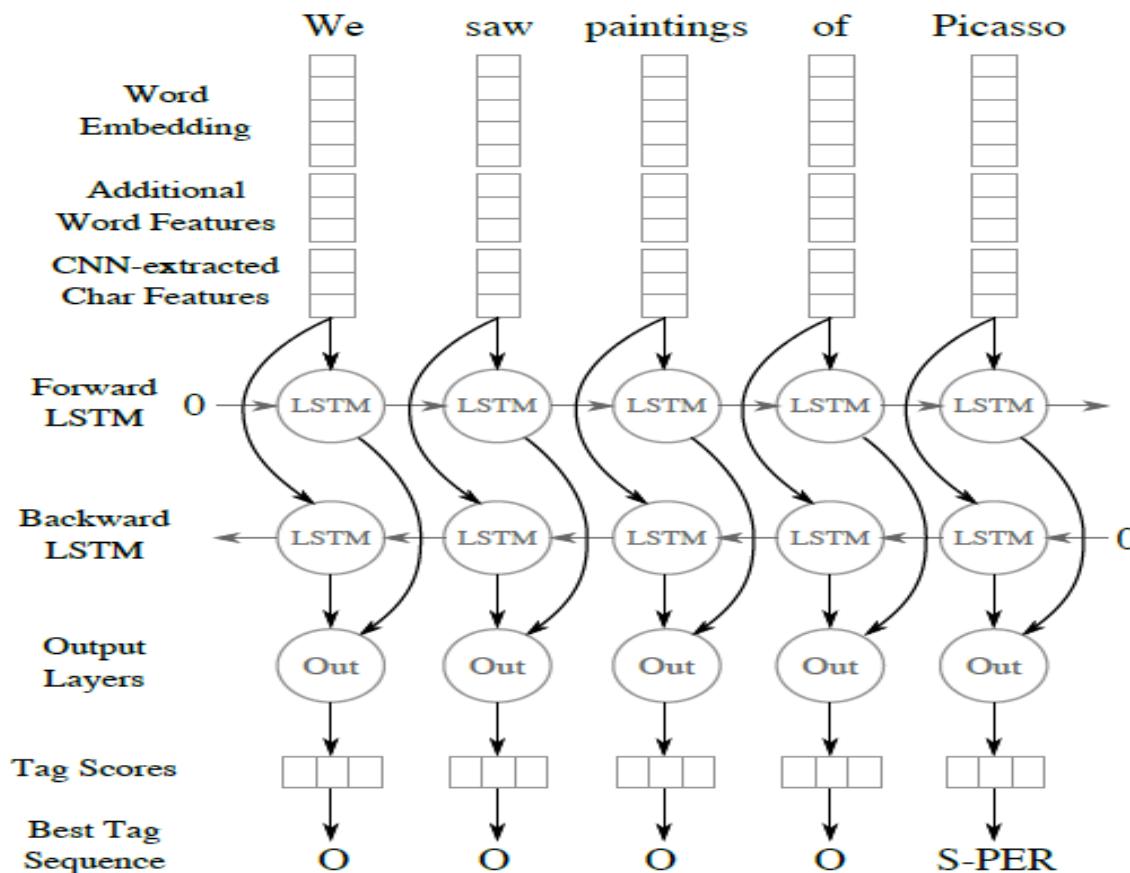
RNNs au niveau caractères:



Approches neuronales pour l'Extraction d'Entités Nommées

CNNs au niveau caractères + LSTMs au niveau mots:

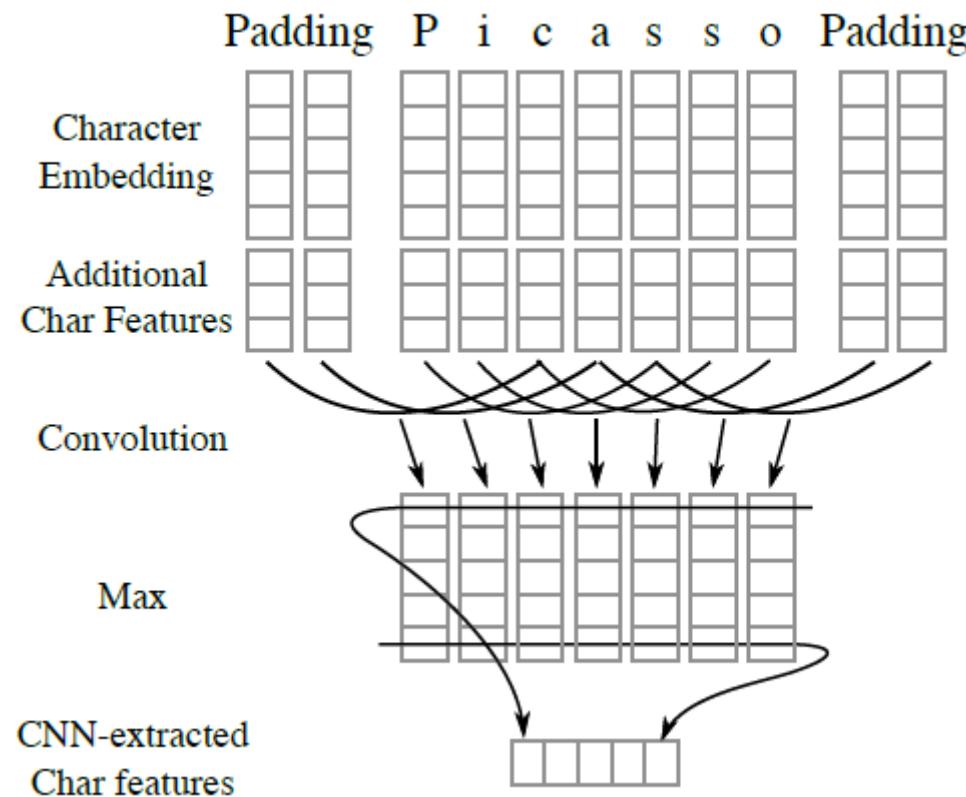
Pour chaque mot les vecteurs au niveau des caractères sont concaténés avec les vecteurs au niveau des mots



Approches neuronales pour l'Extraction d'Entités Nommées

CNNs au niveau caractères:

Le CNN extrait les features de caractères à partir de chaque mot



Analyse sémantique

Le niveau sémantique

- ▶ **But** : résoudre les problèmes de *référence* ; obtenir une *représentation conceptuelle* de l'énoncé dans un langage formel (formules de la logique du premier ordre, graphes conceptuels) ; *articuler* cette représentation conceptuelle avec le monde « physique » de la scène ;
- ▶ **Moyen** : calcul sémantique couplé à l'analyse syntaxique ou traduction ex-post de la représentation arborée dans un langage formel
- ▶ **Outils** : une description sémantique au niveau lexical (relations de synonymie, méronymie, hyper/hyponymie, etc), des règles de composition, des outils de représentation du monde physique ;
- ▶ **Difficultés** : explicitation partielle de l'*implicite* (problèmes de co-référence) ; ambiguïtés sémantiques (portée des quantificateurs) ; taille et précision de la connaissance nécessaire ; choix du formalisme de représentation (temporalité, croyances, etc).
- ▶ **Résultat** : un ensemble de représentations formelles de la scène dans lesquelles les objets et les relations qu'ils entretiennent sont identifiés ;

Le traitement sémantique : résultat

L'arbre syntaxique permet directement d'extraire les propositions (1) à (5), dont on peut déduire, compte-tenu d'une représentation du sens commun, (6), (7), (8) et (9) :

- ▶ $\exists X, \text{president}(X)$: il existe une entité X qui est président (et dont le référent est déjà connu) ;
- ▶ $\exists Y, \text{pomme}(Y)$: il existe une entité Y qui est une pomme ;
- ▶ $\exists Z, \text{couteau}(Z)$: il existe une entité Z qui est un couteau ;
- ▶ $\text{manger}(X, Y)$: cette entité X mange Y ;
- ▶ $\text{moyen}(\text{manger}(X, Y), Z)$: l'opération de manger s'effectue au moyen de Z ;
- ▶ $\text{president}(X) \Rightarrow \text{humain}(X) \Rightarrow \dots$;
- ▶ $(\text{pomme}(Y) \wedge \text{manger}(X, Y)) \Rightarrow \text{aliment}(Y)$
- ▶ $(\text{pomme}(Y) \wedge \text{manger}(X, Y)) \Rightarrow (\text{golden}(X) \mid \text{granny}(X) \mid \dots)$;
- ▶ $\text{manger}(X, Y) \Rightarrow \text{manger}(X), \text{est_ingere}(Y)$;

Chez l'humain, ces déduction se font de manière inconsciente et quasi-réflexe.

Analyse pragmatique

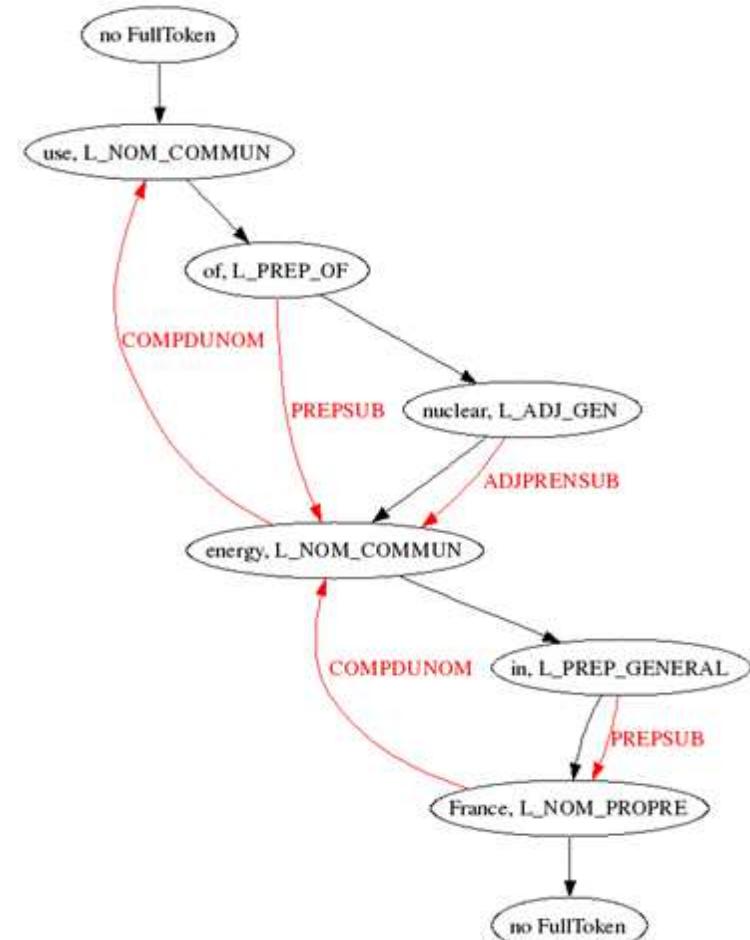
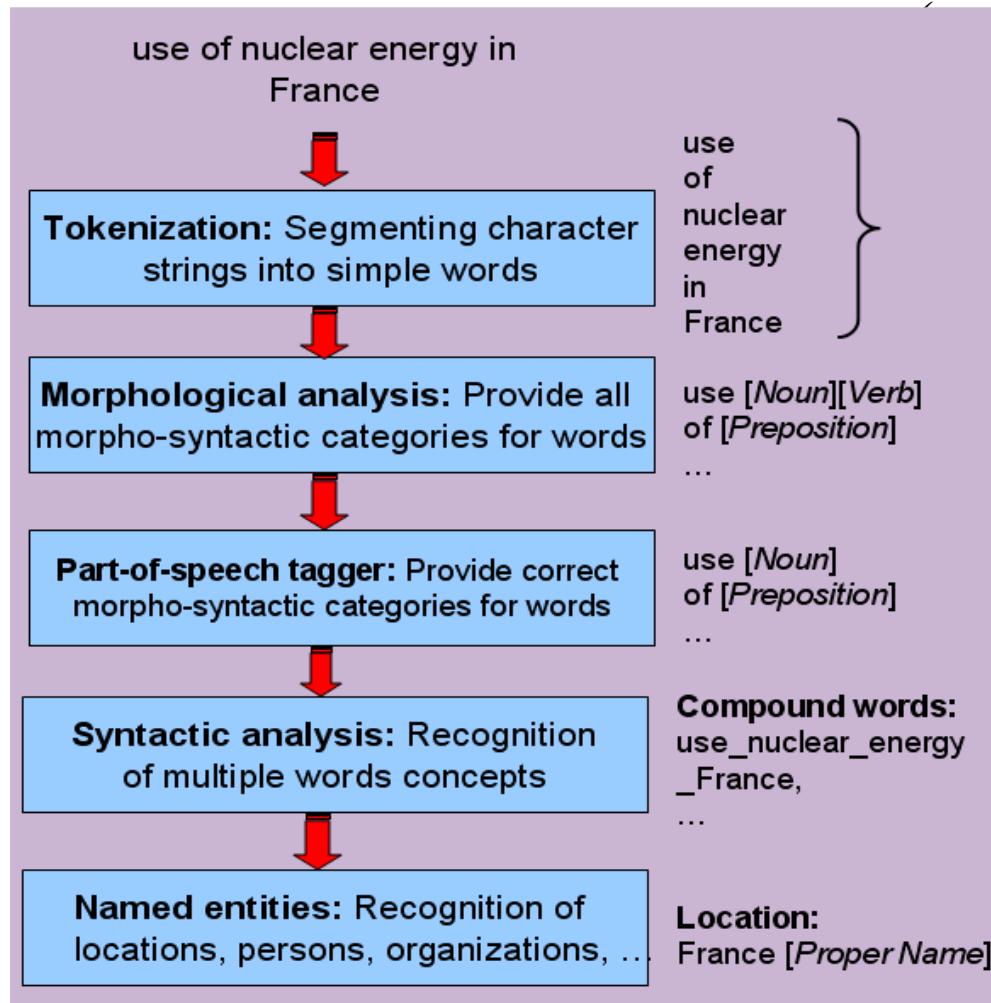
Le niveau pragmatique

- ▶ **But** :achever la *désambiguïsation* de l'énoncé en prenant en compte la dynamique de l'interaction (ou narration) en y intégrant ce qui est implicite ; comprendre la *fonction argumentative* de l'énoncé dans le contexte plus général de l'interaction (ou de la narration) : quelle information nouvelle apporte-t-il, au sujet de quoi dit-il quelque chose, sous quel mode...
- ▶ **Moyen** : une théorie des activités humaines ; une théorie des interactions langagières (la pertinence, les conditions de félicité) ; une théorie des structures discursives...
- ▶ **Outils** : représentation des actions humaines (scripts), « grammaire » des interactions, logique
- ▶ **Difficultés** : taille de la connaissance à représenter, spécification de la « grammaire » des interactions
- ▶ **Résultat** : une représentation formelle contextualisée de l'énoncé, une connaissance de sa fonction argumentative, des connaissances nouvelles...

**Exemple d'un pipeline d'analyse
linguistique:**

LIMA: CEA LIST Multilingual Analyzer

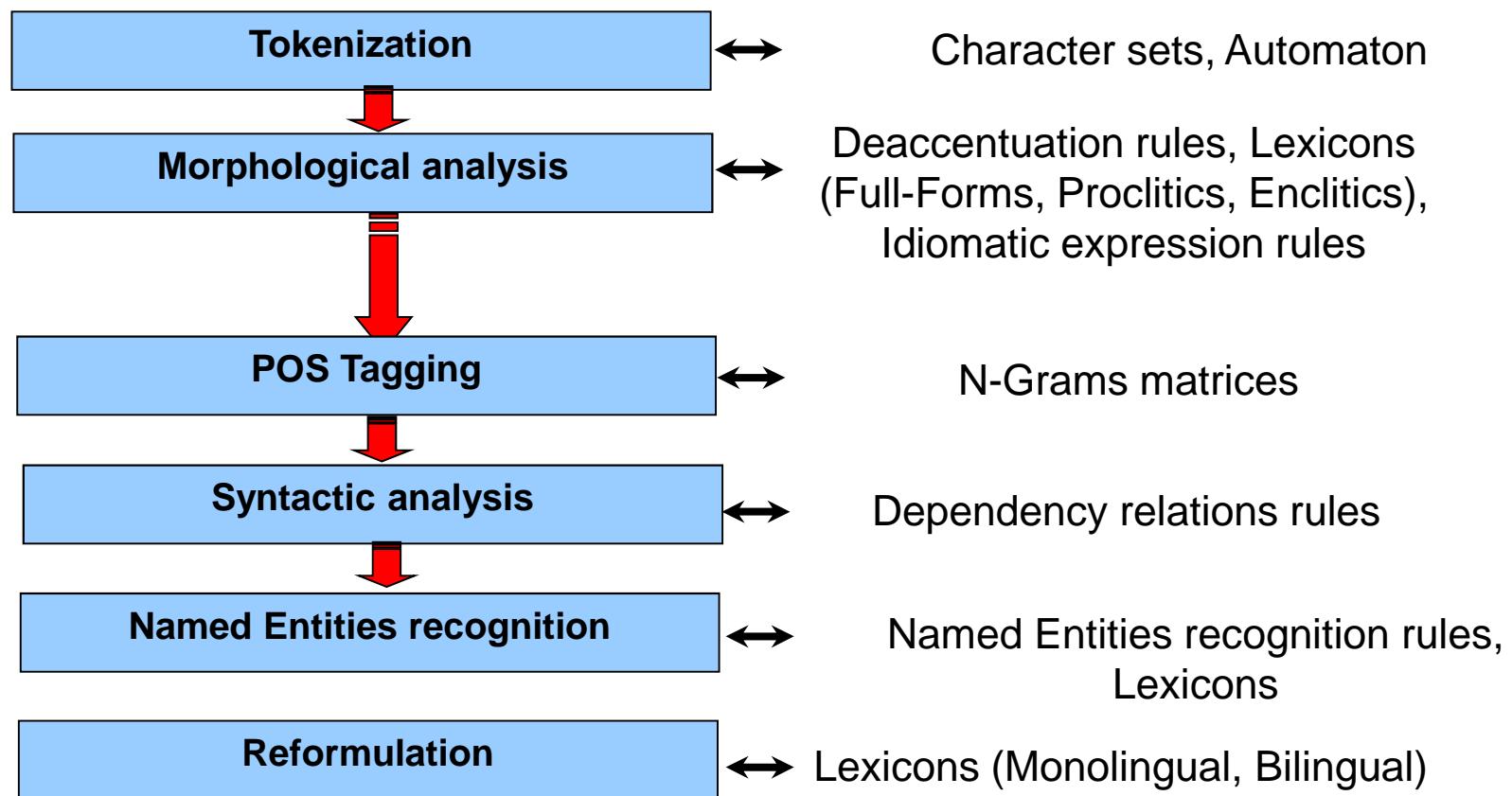
Linguistic Processing



Linguistic processing steps in Information retrieval

- **Morphological analysis**
 - **Tokenization:** Separating the input stream into a graph of words
 - **Simple word lookup:** Search for words in a full-forms lexicon
 - **Orthographical alternative lookup:** Differently accented forms, alternative hyphenisation, concatenated words, abbreviation recognition
 - **Idiomatic expressions recognizer:** Detecting and considering them as single words in the word graph
 - **Critic Stemmer [Only for agglutinated languages]:** Segment the input words into proclitics, simple forms and enclitics
 - **Unknown words analysis**
- **Part-of-Speech tagging and Syntactic analysis**
 - **Part-of-speech tagging:** Using language models from a hand-tagged corpus
 - **Named entity recognizer**
 - **Recognition of nominal and verbal chains in the graph**
 - **Dependency relation extraction**
- **Information retrieval application**
 - **Words graph indexing**
 - **Query reformulation:** Monolingual reformulation for paraphrases and synonymy, multilingual for cross language information retrieval
 - **Retrieval scoring comparing partial matches on subgraphs and entities**

Linguistic resources



Demo

- CEA LIST Multilingual Analyzer (LIMA) – English

- Open source: <https://github.com/aymara/lima/wiki>
- Demo: <http://www.kalisteo.eu/demo/lima/>

LIMA: Multilingual Analyzer

Write or copy text for analysis

The flood of heroin from Asia, cocaine from South America, cannabis from North Africa and synthetic drugs from European bases is unstoppable.

359 Maximum.
Selected Language : English
Detected Coding : utf-8

Basic Options

Language : English

Output Format : Text

Advanced Options

Analyze

About



LIMA

LIMA is the **multilingual analyzer** of the lab. This demo allows you to cut and paste text or submit a file for linguistic analysis.

LIMA's options

The **basic options** let you select the language or the dynamic "autodetection mode" and the output format (text, xml...).

After the analysis, an additional **graph-based representation** is provided.

Advanced options offer more interaction with the system (for advanced users only with skills on *Natural Language Processing*...).

Information

Contact for LIMA
Gael de Chalendar
Mail :gael.de-chalendar{at}cea.fr

Demo

- CEA LIST Multilingual Analyzer (LIMA) – Named Entity Extraction

LIMA	Modify	Reset
<pre><specific_entity> <string>Asia</string> <position>26</position> <length>4</length> <type>Location.LOCATION</type> </specific_entity> <specific_entity> <string>South America</string> <position>45</position> <length>13</length> <type>Location.LOCATION</type> </specific_entity> <specific_entity> <string>North Africa</string> <position>74</position> <length>12</length> <type>Location.LOCATION</type> </specific_entity></pre>	<p>Text</p> <p>The flood of heroin from Asia, cocaine from South America, cannabis from North Africa and synthetic drugs from European bases is unstoppable.</p>	<p>Results</p> <p>1 The the#L_DETERMINANT 5 flood flood#L_NOM_COMMUN 11 of of#L_PREP_GENERAL 14 heroin heroin#L_NOM_COMMUN 21 from from#L_PREP_GENERAL 26 Asia Asia#L_NOM_PROPRE 30 , ,#L_PONCTU_FAIBLE 32 cocaine cocaine#L_NOM_COMMUN 40 from from#L_PREP_GENERAL 45 South America South America#L_NOM_PROPRE 58 , ,#L_PONCTU_FAIBLE 60 cannabis cannabis#L_NOM_COMMUN 69 from from#L_PREP_GENERAL 74 North Africa North Africa#L_NOM_PROPRE 87 and and#L_CONJ_COORD 91 synthetic synthetic#L_ADJ_GEN 101 drugs drug#L_NOM_COMMUN 107 from from#L_PREP_GENERAL 112 European European#L_ADJ_GEN 121 bases base#L_NOM_COMMUN basis#L_NOM_COMMUN 127 is be#L_IS 130 unstoppable unstoppable#L_ADJ_GEN 141 . .#L_PONCTU_FORTE</p>

Demo

- CEA LIST Multilingual Analyzer (LIMA) – French

The screenshot displays the LIMA Multilingual Analyzer interface. On the left, the main workspace shows a text input field containing the French sentence: "En l'absence de ce droit, les demandeurs d'asile pourraient être déplacés d'un État membre à l'autre." Below the text, status information is displayed: "399 Maximum.", "Selected Language : Arabic", and "Detected Coding : Iso8859-1". Under "Basic Options", there are dropdown menus for "Language : French" and "Output Format : Text". A "Advanced Options" button is visible. On the right, three panels provide additional information: "About" (describing LIMA as the multilingual analyzer of the lab, allowing text submission for analysis), "LIMA's options" (explaining basic options like language selection and output format, and graph-based representation), and "Information" (providing contact details for Gael de Chalendar).

LIMA: Multilingual Analyzer

Write or copy text for analysis

En l'absence de ce droit, les demandeurs d'asile pourraient être déplacés d'un État membre à l'autre.

399 Maximum.
Selected Language : Arabic
Detected Coding : Iso8859-1

Basic Options

Language : French

Output Format : Text

Advanced Options

Analyze

About

LIMA

LIMA is the **multilingual analyzer** of the lab. This demo allows you to cut and paste text or submit a file for linguistic analysis.

LIMA's options

The **basic options** let you select the language or the dynamic "autodetection mode" and the output format (text, xml...). After the analysis, an additional **graph-based representation** is provided.

Advanced options offer more interaction with the system (for advanced users only with skills on *Natural Language Processing*...)

Information

Contact for LIMA
Gael de Chalendar
Mail : gael.de-chalendar(at)cea.fr

Demo

- CEA LIST Multilingual Analyzer (LIMA) – POS Tagging

The screenshot shows the LIMA POS Tagging interface. At the top, there's a header with the LIMA logo, a 'Modify' button, and a 'Reset' button. Below the header, the interface is divided into two main sections: 'Text' on the left and 'Results' on the right.

In the 'Text' section, the input text is:

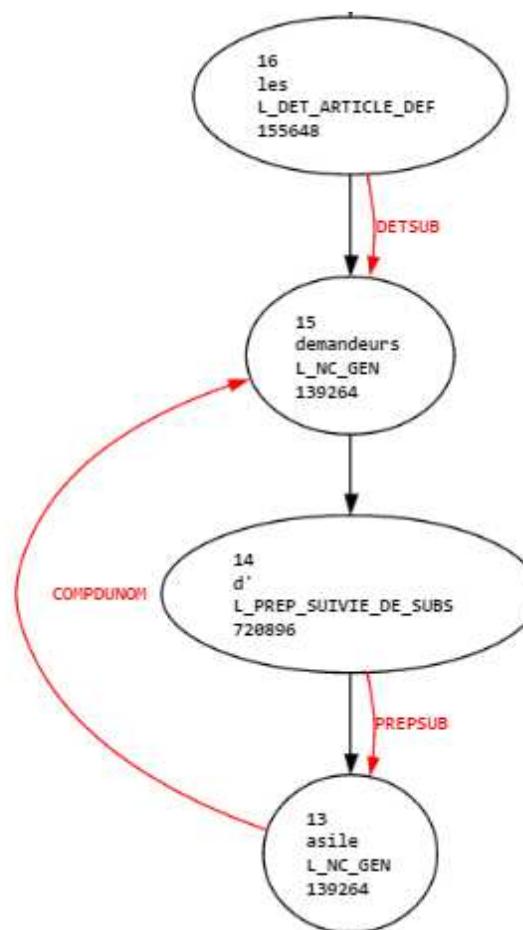
```
En l'absence de ce droit, les demandeurs  
d'asile pourraient être déplacés d'un État  
membre à l'autre.
```

In the 'Results' section, the output is a list of tokens with their corresponding POS tags:

Index	Token	POS Tag
1	En	en#L_PREP_GENERAL
4	l'	le#L_DET_ARTICLE_DEF
6	absence	absence#L_NC_GEN
14	de	de#L_PREP_GENERAL
17	ce	ce#L_DET_DEMONSTRATIF
20	droit	droit#L_NC_GEN
25	,	,#L_PONCTU_FAIBLE
27	les	le#L_DET_ARTICLE_DEF
31	demandeurs	demandeur#L_NC_GEN
42	d'	de#L_PREP_SUIVIE_DE_SUBS
44	asile	asile#L_NC_GEN
50	pourraient	pouvoir#L_VERBE_MODALITE_INDICATIF
61	être	être#L_VERBE_COPULE_INFINITIF
66	déplacés	déplacer#L_VERBE_PRINCIPAL_PARTICIPE_ATTRIB

Demo

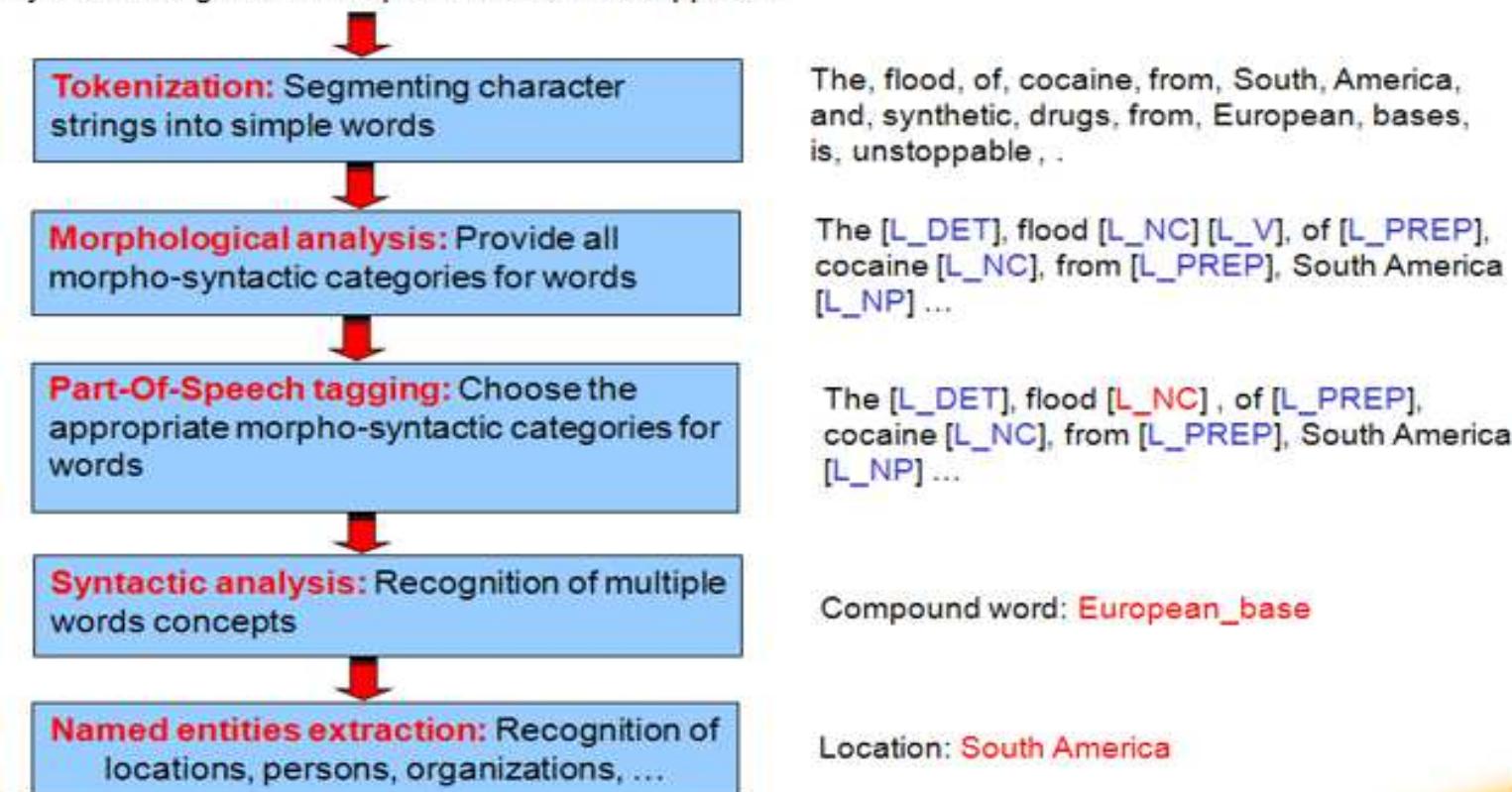
- CEA LIST Multilingual Analyzer (LIMA) – Dependency Parsing



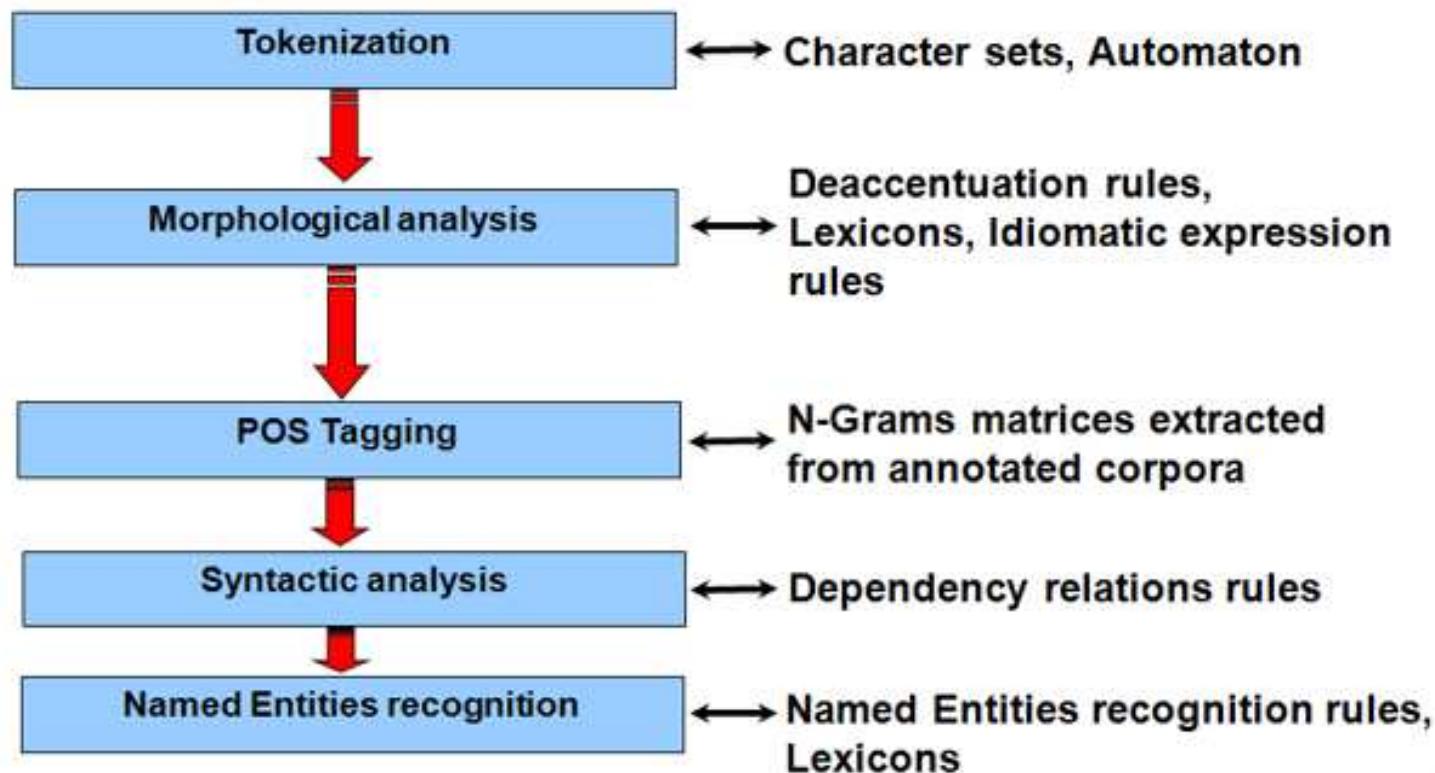
Extraction d'information dans LIMA

Natural Language Processing steps - Example

"The flood of cocaine from South America and synthetic drugs from European bases is unstoppable."



Natural Language Processing - Linguistic resources



LIMA Part-Of-Speech Tagger - Results

The screenshot shows the LIMA POS Tagger application window. On the left, there is an XML input pane containing three specific_entity definitions:

```
<specific_entity>
  <string>Asia</string>
  <position>26</position>
  <length>4</length>
  <type>Location.LOCATION</type>
/<specific_entity>

<specific_entity>
  <string>South America</string>
  <position>45</position>
  <length>13</length>
  <type>Location.LOCATION</type>
/<specific_entity>

<specific_entity>
  <string>North Africa</string>
  <position>74</position>
  <length>12</length>
  <type>Location.LOCATION</type>
/<specific_entity>
```

In the center, there is a text area containing a summary sentence:

The flood of heroin from Asia, cocaine from South America, cannabis from North Africa and synthetic drugs from European bases is unstoppable.

On the right, the results pane displays the POS tags for each word in the sentence. Red boxes highlight several entities: "Asia" at position 26, "South America" at position 45, and "North Africa" at position 74. The results are as follows:

Index	Text	POS Tag
1	The	the\$L_DETERMINANT
5	flood	flood\$L_NOM_COMMUN
11	of	of\$L_PREP_GENERAL
14	heroin	heroin\$L_NOM_COMMUN
21	from	from\$L_PREP_GENERAL
26	(Asia)	Asia\$L_NOM_PROPRE
30	,	,
32	cocaine	cocaine\$L_NOM_COMMUN
40	from	from\$L_PREP_GENERAL
45	(South America)	South America\$L_NOM_PROPRE
58	,	,
60	cannabis	cannabis\$L_NOM_COMMUN
69	from	from\$L_PREP_GENERAL
74	(North Africa)	North Africa\$L_NOM_PROPRE
87	and	and\$L_CONJ_COORD
91	synthetic	synthetic\$L_ADJ_GEN
101	drugs	drug\$L_NOM_COMMUN
107	from	from\$L_PREP_GENERAL
112	European	European\$L_ADJ_GEN
121	bases	base\$L_NOM_COMMUN
127	is	be\$L_IS
130	unstoppable	unstoppable\$L_ADJ_GEN
141	.	,

LIMA Information Extraction tasks

- **Named Entity Recognition (NER)**
 - Identification of specific entities in the texts
 - Co-reference
- **Relation extraction**
 - Identification of relation existing between two entities
 - Attributive (e.g. date-of-birth)
 - Event-related relations (e.g. acquisition)
- **Event Extraction**
 - Identification of a given set of relations tied by a template structure

LIMA Named Entity Recognition - Example

Bob Boulton, born in 1950, has been appointed President of Minelco, effective December 1, 2010. He succeeds Markus Petäjäniemi, who has been appointed LKAB's Senior Vice President.

Entities

PERSON

DATE

FUNCTION

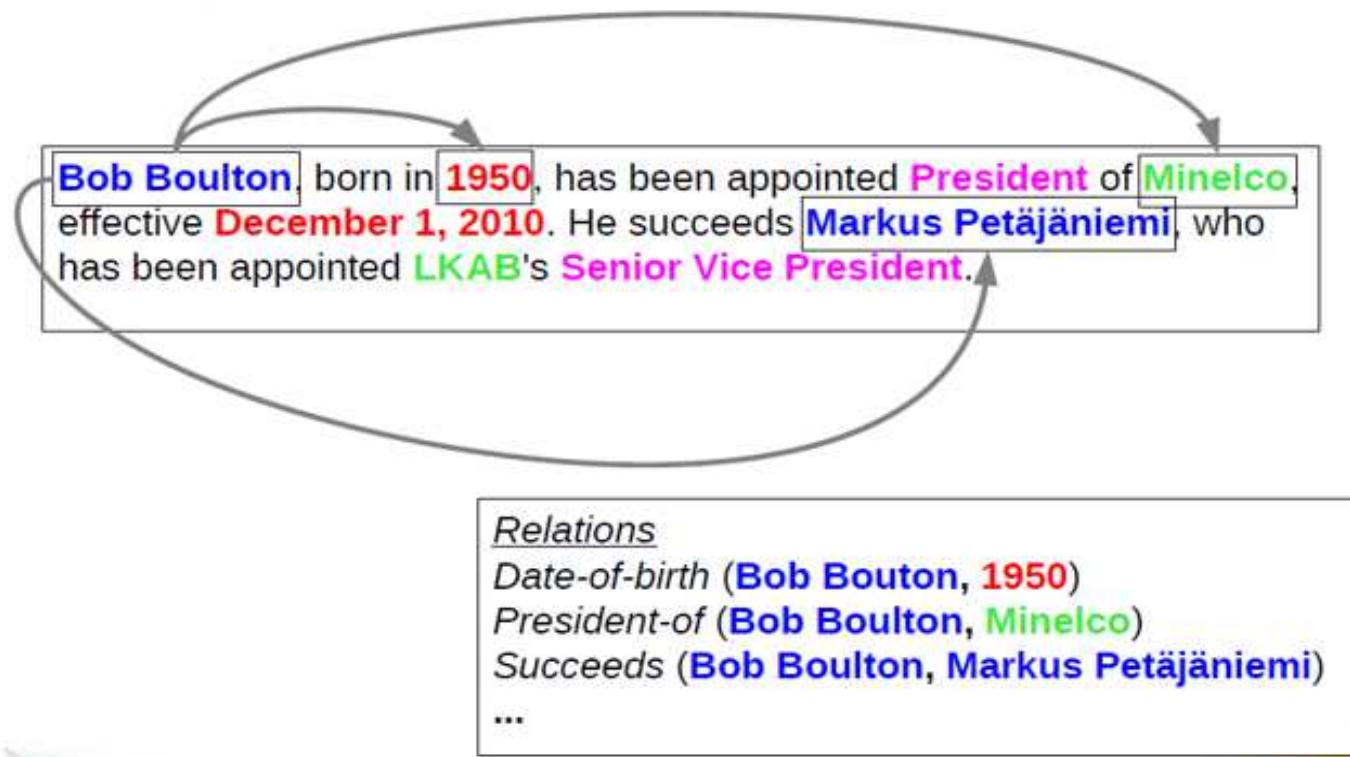
ORGANIZATION

Co-reference

He

→ Bob Boulton

LIMA Relation Extraction - Example



LIMA Rule-based Named Entity Recognition

- Rule format

```
<trigger> : <preceding context> : <following context> : <type> : =><action>
```

- regular expressions based on linguistic information

- Examples for entity LOCATION

```
@RegionsAndCountries=(Afghanistan,Africa,Albania,Alberta,...)  
@City=(Aaccra,Aalborg,Aarhus,Ababa,Abadan,Abakan,Aberdeen,...)  
@GeographicalPrecision=(South,West,North,East,south,west,north,east,Southern,  
Western,Northern,Eastern,southern,western,northern,eastern)  
@locationKey=(sea,ocean,Village,Valley,Trail,Station,Stadium,Square,River,...)  
  
@GeographicalPrecision:::(@City|@CountryOrRegion):LOCATION:  
  
$L_NP:(@GeographicalPrecision)?:@locationKey:LOCATION:
```

LIMA Rule-based Named Entity Recognition

■ Example of rule composition from annotated text

Bob Boulton, born in **1950**, has been appointed **President** of **Mineco**, effective **December 1, 2010**. He succeeds **Markus Petäjäniemi**, who has been appointed **LKAB's Senior Vice President**.

```
@function=(President, Director,...)  
@functionModifier=(Senior, Junior, Executive)  
@functionModifier2=(Vice)  
@function:@functionModifier{0-2} @functionModifier2? @function::FUNCTION:  
[<FUNCTION>]::[of] $PROPER_NOUN ($PROPER_NOUN){0-2}:ORGANIZATION:  
[<FUNCTION>]:$PROPER_NOUN ($PROPER_NOUN){0-2} ['s]:ORGANIZATION:
```

Outils de TALN

Traitement Automatique de la Langue

- Natural Language Software Registry : <http://registry.dfki.de>
- OpenNLP : <http://opennlp.sourceforge.net/projects.html>
- Standford NLP : <http://nlp.stanford.edu/software>
- CCG Software : <http://cogcomp.cs.illinois.edu/page/software>
- CLARIN : http://www.clarin.eu/view_tools
- ATALA : <http://www.atala.org/-Outils-pour-le-TAL->
- IMS : <http://www.ims.uni-stuttgart.de/projekte/gramotron/resources.html>
- ISLanD : <https://www.greyc.fr/node/8?q=node/884>

Extraction d'Entités Nommées

- HeidelTime (expressions temporelles)
<http://dbs.ifi.uni-heidelberg.de/index.php?id=129>
- GeniaTagger (entités nommées en génomiques)
<http://www.nactem.ac.uk/GENIA/tagger/>
- LIA NE (pour le français)
<http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html>
- Stanford NER (pour l'anglais, l'allemand, l'espagnol et le chinois) <http://nlp.stanford.edu/ner/index.shtml>