Traitement Automatique des Langues
Laboratoire d'Informatique Fondamentale
équipe TALEP

Alexis Nasr

November 23, 2017

Activités de l'équipe

- Développement d'outils et de ressources génériques pour le Traitement Automatique des Langues
  - analyseurs morphologiques
  - étiqueteurs morpho-syntaxiques
  - analyseurs syntaxiques
  - détecteurs d'entités nommées
  - analyseurs sémantiques
  - analyseurs discursifs
- Utilisation dans le cadre d'applications
  - extraction d'information
  - analyse de discours
  - résumé de textes
  - détection/correction d'erreurs

Traitement Automatique des Langues

- Tentative de reproduire avec des ordinateurs certains traitements linguistiques réalisés par l'homme
- Vision boîte noire : pour une "entrée" X, on essaye de reproduire les "sorties" Y
- On n'essaye pas de reproduire le fonctionnement du cerveau humain
- On essaye de tirer le meilleur parti de l'ordinateur
- Les modèles sont conçus pour minimiser les erreurs sur la tâche qu'ils doivent résoudre
  - Ils nous apprennent certaines choses sur la langue
  - Ils ne nous apprennent rien sur le traitement de la langue par le cerveau

Traduction

- Computational Linguistics
- Natural Language Processing

Quelques interfaces

- Informatique
  - Algorithmique
  - Théorie des langages
  - Apprentissage automatique
  - Calculabilité
  - Complexité
  - ...
- Linguistique
  - Linguistique formelle
  - Linguistique de corpus
  - Linguistique expérimentale
  - ...
- Psychologie
- Neurosciences

Quelques interfaces

- Informatique
  - Algorithmique
  - Théorie des langages
  - Apprentissage automatique
  - Calculabilité
  - Complexité
  - ...
- Linguistique ♡
  - Linguistique formelle
  - Linguistique de corpus
  - Linguistique expérimentale
  - ...
- Psychologie
- Neurosciences

Outils

Structuration du texte ou du signal de parole

Trois opérations formelles :

- Segmentation
  - d'un texte en phrases
  - d'un tour de parole en unités macrosyntaxiques
  - d'une phrase en mots
  - d'une phrase en entités nommées
  - d'un mot en morphèmes
- Etiquetage
  - d'un mot en partie de discours
  - d'une conversation en actes de dialogues
  - d'un document en thèmes
  - d'un mot en sens
- Etablissement de relations
  - morphologiques
  - syntaxiques
  - sémantiques
  - discursives

Architecture générique (version naïve)

Etant donné une entrée X

1. Enumération de toutes les solutions possibles

$$\mathcal{Y} = \{Y_1 \dots Y_n\}$$

2. Pondération des solutions

$$p(Y_i)$$

3. Sélection de la solution de meilleur poids

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} p(Y)$$

Enumération : Segmentation

$X =$	a	b	c	d

$\mathcal{Y} =$				
	a	b	c	d
	a	b	c	d
	a	b	c	d
	a	b	c	d
		...		
	a	b	c	d

- 2^{n-1} segmentations possibles

Enumération : Etiquetage

$X =$	a	b	c	d
	1	1	1	1
	1	1	1	2
$\mathcal{Y} =$	1	1	2	1
		...		
	k	k	k	k

- k^n étiquetages possibles
- la segmentation peut être vue comme un cas particulier d'étiquetage (2 étiquettes : $\{$`debut`, `interieur`$\}$)

Enumeration : Relations

$X =$	a	b	c	d
	a	a	a	a
	a	a	a	b
$\mathcal{Y} =$	a	a	b	a
		...		
	d	d	d	d

- Relation *est le fils de*
- n^n graphes possibles

Pondération

- Le poids d'une solution est une fonction des poids des parties : (unités minimales)

$$p(Y) = \sum_{y \in \mathcal{F}(Y)} p(y)$$

- $p(y)$ est le poids de la partie y
- La décomposition de Y en parties est un compromis :
  - trop petites, elles ne sont pas très riches linguistiquement
  - trop grosses, leur poids est difficile à estimer

Estimation des poids

Les poids des parties sont estimés à partir de corpus annotés (X_i, Y_i)

- Modèles génératifs

$$\hat{M} = \arg\max_M P_M(X, Y)$$

- Modèles discriminants

$$\hat{M} = \arg\max_M P_M(Y|X)$$

Recherche de la solution de meilleur poids

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} \sum_{y \in Y} p(y)$$

Recherche de la solution de meilleur poids

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} \sum_{y \in Y} p(y)$$

En pratique $\mathcal{Y}$ n'est jamais énuméré

- Ce n'est pas envisageable d'un point de vue calculatoire
- Ce n'est pas satisfaisant car la plupart des solutions partagent des parties communes
- L'ensemble des solutions est représenté sous la forme d'une structure partagée $\mathcal{F}(\mathcal{Y})$

$$\hat{Y} = \arg\max_{Y \in \mathcal{F}(\mathcal{Y})} \sum_{y \in Y} p(y)$$

Quelques adjectifs

- séquentiels
- combinatoires
- empiriques
- statiques
- non introspectifs

Séquentiels

- La plupart des tâches mettent en jeux plusieurs processus (modules)
- Exemple :
  - segmentation en phrases
  - segmentation en mots
  - étiquetage morpho-syntaxique
  - analyse syntaxique
- L'espace de recherche global est trop gros pour être représenté
- Les processus sont généralement organisés de manière séquentielle
- Certains choix sont effectués prématurément
  - *Je mange bien que je n'aie pas faim*
  - *Je pense bien que je l'ai oublié*

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	faim

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	faim
je	mange	bien que		je	n'	aie	pas	faim

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	faim
je	mange	bien que		je	n'	aie	pas	faim
CLI	VRB	CSU		CLI	ADN	VRB	ADV	NOM

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	faim
je	mange	bien que		je	n'	aie	pas	faim
CLI	VRB	CSU		CLI	ADN	VRB	ADV	NOM
SUJ 2	ROOT	MOD 2		SUJ 7	MOD 7	OBJ 3	MOD 7	OBJ 7

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je								

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je								
je								

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je								
je								
CLI								

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je								
je								
CLI								
?								

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je								
je								
CLI								
?								

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange							
je								
CLI								
?								

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange							
je	mange							
CLI								
?								

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange							
je	mange							
CLI	VRB							
?								

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange							
je	mange							
CLI	VRB							
SUJ	ROOT							

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange							
je	mange							
CLI	VRB							
SUJ	ROOT							

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien						
je	mange							
CLI	VRB							
SUJ	ROOT							

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien						
je	mange	bien						
CLI	VRB							
SUJ	ROOT							

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien						
je	mange	bien						
CLI	VRB	ADV						
SUJ	ROOT							

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien						
je	mange	bien						
CLI	VRB	ADV						
SUJ	ROOT	MOD						

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien						
je	mange	bien						
CLI	VRB	ADV						
SUJ	ROOT	MOD						

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que					
je	mange	bien						
CLI	VRB	ADV						
SUJ	ROOT	MOD						

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que					
je	mange	bien que						
CLI	VRB							
SUJ	ROOT							

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que					
je	mange	bien que						
CLI	VRB	CSU						
SUJ	ROOT							

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que					
je	mange	bien que						
CLI	VRB	CSU						
SUJ	ROOT	MOD						

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que					
je	mange	bien que						
CLI	VRB	CSU						
SUJ	ROOT	MOD						

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je				
je	mange	bien que						
CLI	VRB	CSU						
SUJ	ROOT	MOD						

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je				
je	mange	bien que		je				
CLI	VRB	CSU						
SUJ	ROOT	MOD						

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je				
je	mange	bien que		je				
CLI	VRB	CSU		CLI				
SUJ	ROOT	MOD						

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je				
je	mange	bien que		je				
CLI	VRB	CSU		CLI				
SUJ	ROOT	MOD		?				

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je				
je	mange	bien que		je				
CLI	VRB	CSU		CLI				
SUJ	ROOT	MOD		?				

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'			
je	mange	bien que		je				
CLI	VRB	CSU		CLI				
SUJ	ROOT	MOD		?				

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'			
je	mange	bien que		je	n'			
CLI	VRB	CSU		CLI				
SUJ	ROOT	MOD		?				

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'			
je	mange	bien que		je	n'			
CLI	VRB	CSU		CLI	ADN			
SUJ	ROOT	MOD		?				

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'			
je	mange	bien que		je	n'			
CLI	VRB	CSU		CLI	ADN			
SUJ	ROOT	MOD		?	?			

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'			
je	mange	bien que		je	n'			
CLI	VRB	CSU		CLI	ADN			
SUJ	ROOT	MOD		?	?			

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie		
je	mange	bien que		je	n'			
CLI	VRB	CSU		CLI	ADN			
SUJ	ROOT	MOD		?	?			

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie		
je	mange	bien que		je	n'	aie		
CLI	VRB	CSU		CLI	ADN			
SUJ	ROOT	MOD		?	?			

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie		
je	mange	bien que		je	n'	aie		
CLI	VRB	CSU		CLI	ADN	VRB		
SUJ	ROOT	MOD		?	?			

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie		
je	mange	bien que		je	n'	aie		
CLI	VRB	CSU		CLI	ADN	VRB		
SUJ	ROOT	MOD		SUJ	MOD	OBJ		

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie		
je	mange	bien que		je	n'	aie		
CLI	VRB	CSU		CLI	ADN	VRB		
SUJ	ROOT	MOD		SUJ	MOD	OBJ		

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	
je	mange	bien que		je	n'	aie		
CLI	VRB	CSU		CLI	ADN	VRB		
SUJ	ROOT	MOD		SUJ	MOD	OBJ		

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	
je	mange	bien que		je	n'	aie	pas	
CLI	VRB	CSU		CLI	ADN	VRB		
SUJ	ROOT	MOD		SUJ	MOD	OBJ		

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	
je	mange	bien que		je	n'	aie	pas	
CLI	VRB	CSU		CLI	ADN	VRB	ADV	
SUJ	ROOT	MOD		SUJ	MOD	OBJ		

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	
je	mange	bien que		je	n'	aie	pas	
CLI	VRB	CSU		CLI	ADN	VRB	ADV	
SUJ	ROOT	MOD		SUJ	MOD	OBJ	MOD	

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	
je	mange	bien que		je	n'	aie	pas	
CLI	VRB	CSU		CLI	ADN	VRB	ADV	
SUJ	ROOT	MOD		SUJ	MOD	OBJ	MOD	

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	faim
je	mange	bien que		je	n'	aie	pas	
CLI	VRB	CSU		CLI	ADN	VRB	ADV	
SUJ	ROOT	MOD		SUJ	MOD	OBJ	MOD	

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	faim
je	mange	bien que		je	n'	aie	pas	faim
CLI	VRB	CSU		CLI	ADN	VRB	ADV	
SUJ	ROOT	MOD		SUJ	MOD	OBJ	MOD	

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	faim
je	mange	bien que		je	n'	aie	pas	faim
CLI	VRB	CSU		CLI	ADN	VRB	ADV	NOM
SUJ	ROOT	MOD		SUJ	MOD	OBJ	MOD	

je mange bien que je n'aie pas faim

1	2	3	4	5	6	7	8	9
je	mange	bien	que	je	n'	aie	pas	faim
je	mange	bien que		je	n'	aie	pas	faim
CLI	VRB	CSU		CLI	ADN	VRB	ADV	NOM
SUJ	ROOT	MOD		SUJ	MOD	OBJ	MOD	OBJ

Combinatoires

- Toutes les solutions sont envisagées
- Les plus farfelues sont écartées du fait de leur poids
- Prix de l'agnosticisme linguistique
- On pourrait utiliser :
  - une grammaire générative
  - des contraintes linguistiques
  - des contraintes cognitives
  - des contraintes neurologiques ?

Empiriques

- Les modèles sont appris sur des données (généralement annotées)
- Ce qui n'existe pas dans les données n'est pas modélisé
- Modèle de la performance et non de la compétence (variabilité)

Statiques

- L'ordinateur apprend vite
- Deux phases : Apprentissage, Utilisation
- Après l'étape d'apprentissage, l'ordinateur n'apprend plus

Non conscients

- L'ordinateur n'a généralement pas conscience de ses erreurs
- Mesures de confiance

Quelques interfaces

- Informatique
- Linguistique
- Psychologie
  - Fournir des outils de TAL (simplification)
  - Veut on essayer de se rapprocher des comportements humains ?
- Neurosciences
  - Que peuvent apporter les neurosciences au TAL ?
  - Que peut apporter le TAL aux neurosciences ?