Le Cadre Général

Traitements

Une grande partie des tâches de TAL peuvent être vues comme la 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
	a |	b |	c |	d
	a	b |	c |	d
$\mathcal{Y} =$	a |	b	c |	d
	a |	b |	c	d
		$\dots$		
	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 grandes, 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 conscients

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*

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
- On peut cependant définir des mesures de confiance, qui permettent de modéliser la confiance qu'à l'ordinateur dans la solution qu'il propose.