Segmentation

Béatrice Daille - Université de Nantes, LINA

19 septembre 2013

- ⋆ **Problématique**
- ⋆ **Segmentation en phrases : algorithmes**
- ⋆ **Reconnaissance unités lexicales**

Problématique

⋆ **segmentation en phrase**

Langues idéographiques (Chinois, Japonais)
Pas de délimiteurs de mots, ni de phrases.
Segmentation étape déterminante pour des traitements plus élaborés.

◦ **Frontières entre mots**

Chinois : taux de reconnaissance des mots est de 76 % entre locuteurs natifs

⋆ **Reconnaissance unités lexicales**

Langues utilisant des caractères séparateur de mots
phrasèmes
composés

Segmentation en phrases

Ambiguïtés

⋆ **Ambiguïté croisée**

Étant donné une chaine, ABC, deux mots possibles : AB et BC.

白天鹅 (cygne blanc) peut être segmenté comme 白+ 天鹅 (blanc + cygne) ou 白天 + 鹅 (jour + oie domestique).

⋆ **Ambiguïté combinatoire**

Étant donné une chaine, AB, trois mots possibles : A, B et AB.

矛盾 (contradiction) peut être segmenté comme 矛盾 ou 矛 + 盾 (lance + bouclier).

◦ **Mots inconnus**

Huang et Zhao (2007) ont montré que la perte de la précision due aux mots inconnus est 5 fois supérieure à celle causée par l'ambiguïté.

Algorithmes de segmentation

Entrée : **Dictionnaire** + Texte à segmenter

◦ **Recherche maximale en avant (Forward maximum matching FMM)**
parcourt de gauche à droite pour trouver la chaine la plus longue ∈ dictionnaire
英文章鱼怎么说 (Comment dire le poulpe en anglais ?)
FMM : 英文 anglais 章鱼 poulpe 怎么 comment 说 dire
segmentation correcte

◦ **Recherche maximale en arrière (Backward maximum matching BMM)**
parcourt de droite à gauche pour trouver la chaine la plus longue ∈ dictionnaire
英文章鱼怎么说 (Comment dire le poulpe en anglais ?)
BMM : 英 angleterre 文章鱼 article 怎么 comment 说 dire
segmentation incorrecte

◦ **Recherche maximale bidirectionnelle (Bidirection maximum matching BiMM)**
Application de FMM puis BMM. Si différence de segmentation, on prend la segmentation minimale.

Segmenteur ICTCLAS (1)

Algorithme de ICTCLAS : segmentation du chinois état de l'art

Segmentation atomique étape initiale

segmentation en unités lexicales minimales

ilsont

	1	2	3	4	5	6	7	8
0	<début>							
1		i						
2			l					
3				s				
4					o			
5						n		
6							t	
7								<fin>

Segmenteur ICTCLAS (2)

Pré-segmentation Construction d'un graphe orienté de segmentation.

Sur une même ligne : mots qui commencent par la même lettre

Numéro de colonne d'un segment courant = numéro de ligne de son successeur (successeur de la colonne 4 = ligne 4)

	1	2	3	4	5	6	7	8
0	<début>							
1		i	il	ils				
2			l					
3				s		son	sont	
4					o	on	ont	
5						n		
6							t	
7								<fin>

Segmenteur ICTCLAS (3)

N-plus courts chemins Algorithme de Dijkstra

Calcul de la longueur d'un arc : probabilité jointe

$L_{arc} = -log(P(W))$

Calcul de la longueur d'un chemin

$L_{chemin} = -\sum_{i=1}^{n-1} log(P(W_i))$

probabilités estimées avec unigrammes

Segmenteur ICTCLAS (4)

Rajout du contexte

Calcul de la longueur d'un arc

$L_{arc} = -log(P(W_2|W1))$

Calcul de la longueur d'un chemin

$L_{chemin} = -\sum_{i=1}^{n-1} log(P(W_{i+1}|W_i))$

probabilités estimées avec bigrammes

	1	2	3	4	5
0	<début>@il 5,64	<début>@ils 5,65			
1			il@sont 7,75		
2				ils@ont 2,07	
3					sont@<fin> 7,48
4					ont@<fin> 7,35

Évaluation segmentation

Calcul de la distance d'édition

Référence	$C_1C_2SC_3C_4C_5SC_6C_7C_8C_9SC_{10}SC_{11}$
Segmenteur	$C_1C_2SC_3C_4C_5SC_6C_7SC_8C_9SC_{10}C_{11}$