# Approches fondées sur les chaînes de caractères pour le Recherche d'Information

Mathieu Roche

Cours ECD (Recherche d'Information et Langage Naturel)

2008/2009

# Utilisation des informations sur les chaînes de caractères en RI

- **Utiliser des connaissances sémantiques pour améliorer les méthodes de classification (cf cours précédent).**
  - De telles connaissances existent dans le domaine général.
  - *Limite :* domaines spécialisés.

**Lien entre les chaînes de caractères et la “sémantique” ?**

# Utilisation des informations sur les chaînes de caractères en RI

- **Utiliser des méthodes fondées sur les chaînes de caractères pour :**
  - **Apporter des connaissances sémantiques** (pour le regroupement de mots “sémantiquement” proches),
  - **Normaliser les textes** (correction orthographique, etc.),
  - **Reconnaissance des langues**,
  - **Identification de plagiat** (proximité de marques déposées à l'INPI),
  - etc.

# Suffixes/Préfixes

- **But :** vérifier qu'une chaîne de caractères *Ch1* se retrouve :
  - au début d'une chaîne de caractères *Ch2* (**préfixe**),
  - à la fin d'une chaîne de caractères *Ch2* (**suffixe**).

- *Exemples de similarités :*
  - Préfixe -> *Ch1* = chat / *Ch2* = chaton
  - Suffixe -> *Ch1* = suivre / *Ch2* = poursuivre

# Suffixes/Préfixes

- ***Avantage :*** efficace sur certains **domaines spécialisés** tels que la médecine [Nakache *et al.* 2006]

✔ Les suffixes indicateurs d'**états pathologiques** : 'ite' pour désigner l'inflammation (pancréatite, appendicite, gastrite), 'algie' ou 'odynie' pour la douleur.

✔ Les suffixes indicateurs de **gestes techniques** : 'centèse' signifie ponction, 'ectomie' est propre à l'ablation, 'plastie' la réparation.

# Suffixes/Préfixes

- Utilisation de ces connaissances (suffixes/préfixes) sur les chaînes de caractères comme connaissance du domaine.

- Désuffixation pour améliorer les méthodes de classification [Nakache *et al.*, 2006]

- ***Limite :*** chat / chateau !

# String Matching

- Il existe de nombreuses mesures de similarité (pas seulement au niveau des méthodes de mise en correspondance de schémas).

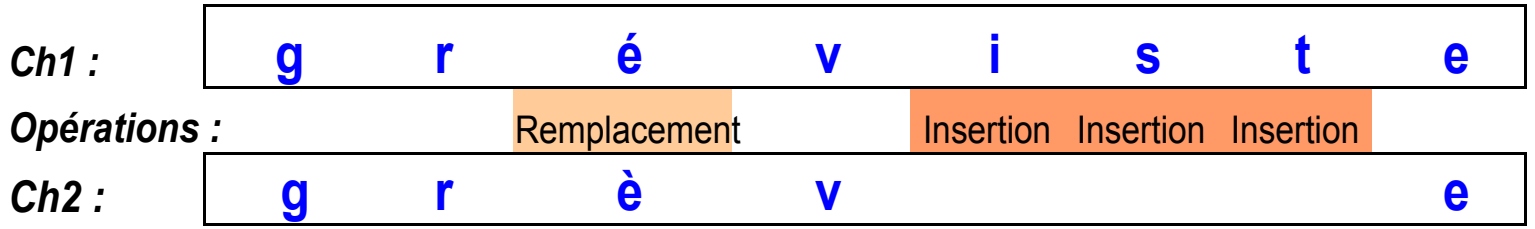
- Exemple avec la distance « Edit distance » (notée *E*) = somme minimale du coût des opérations qu'il faut effectuer pour transformer *Ch1* en *Ch2*.

  *Opérations : suppression, insertion, remplacement.*

**Remarque :** L'« Edit Distance » est aussi appelé « Distance de Levenshtein »

# String Matching

- Exemple : *E*(*gréviste*,*grève*) = 4




Mesure prenant en compte *E* : la mesure String Matching (*SM*) de Maedche et Staab :

$$SM(Ch1,Ch2) = \max[\,0;\ (\min(|Ch1|,|Ch2|) - E(Ch1,Ch2))/\min(|Ch1|,|Ch2|)\,]$$

- *SM*(*gréviste*,*grève*) = max(0;(5-4)/5) = 0.2
- Calculer *SM*(chat,chaton)

# String Matching

**Méthode (Distance de Levenshtein) :**

Construire une matrice M de n+1 lignes et m+1 colonnes. Initialiser de la première ligne par la matrice ligne [ 0,1,….., m-1, m] et la première colonne par la matrice colonne [ 0,1,….., n-1, n]

| | | C | H | I | E | N | S |
|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| N | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| I | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| H | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 5 | 0 | 0 | 0 | 0 | 0 | 0 |

Soit Cout(i, j)=0 si A(i)=B(j) et Cout(i, j)=1 si A(i)!=B(j) On a donc ici la matrice Cout :

| | C | H | I | E | N | S |
|---|---|---|---|---|---|---|
| N | 1 | 1 | 1 | 1 | 0 | 1 |
| I | 1 | 1 | 0 | 1 | 1 | 1 |
| C | 0 | 1 | 1 | 1 | 1 | 1 |
| H | 1 | 0 | 1 | 1 | 1 | 1 |
| E | 1 | 1 | 1 | 0 | 1 | 1 |

# String Matching

On remplit ensuite la matrice M en utilisant la règle suivante  M[i, j] est égale au minimum de:

- ***L'élément directement avant plus 1: M[i-1, j] + 1.***
- ***L'élément directement au dessus plus 1: M[i, j-1] + 1.***
- ***Le diagonal précédent plus le coût: M[i-1, j-1] + Cout(i, j).***

| | | C | H | I | E | N | S |
|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| N | 1 | 1 | 2 | 3 | 4 | 4 | 5 |
| I | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| H | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 5 | 0 | 0 | 0 | 0 | 0 | 0 |

…

| | | C | H | I | E | N | S |
|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| N | 1 | 1 | 2 | 3 | 4 | 4 | 5 |
| I | 2 | 2 | 2 | 2 | 3 | 4 | 5 |
| C | 3 | 2 | 3 | 3 | 3 | 4 | 5 |
| H | 4 | 3 | 2 | 3 | 4 | 4 | 5 |
| E | 5 | 4 | 3 | 3 | 3 | 4 | 5 |

Calculer la matrice pour les mots :

*(chat, chaton)*

# n-grammes

- Technique des *n*-grammes est utilisée pour calculer le nombre de *n* caractères consécutifs.

- Généralement, la valeur de *n* varie entre 1 et 5.

  - *Exemple de tri-grammes : Ch1 = chat / Ch2 = chaton :*
  - *tr(Ch1) = {cha, hat}*
  - *tr(Ch2) = {cha, hat, ato, ton}*

- *Mise en oeuvre de mesures fondées sur les tri-grammes tels que la mesure de Lin.*

# n-grammes

- Mesure *Lin (1998) :*

$$Tri(Ch1,Ch2) = 1/[\ 1 + |tr(Ch1)| + |tr(Ch2)| - 2\ X\ |tr(Ch1)\ Intersect\ tr(Ch2)|\ ]$$

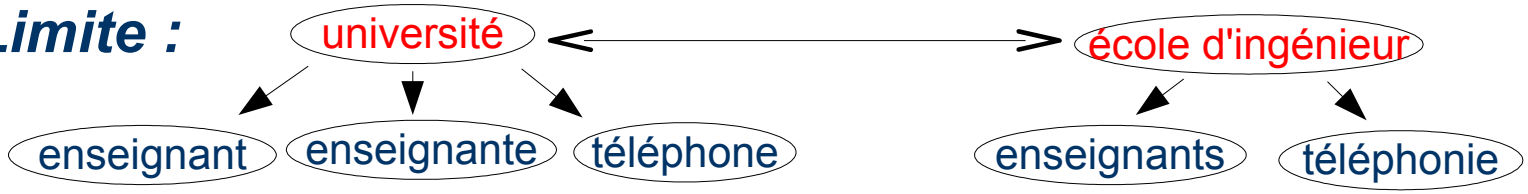
- *SM*(chat,chaton) = 1/[1+2+4-2X2]=0.33

# Avantages et limites des mesures fondées sur les chaînes de caractères

- Mesures fondées sur les chaînes de caractères souvent utilisées pour la mise en correspondances de schémas.
  - Reconnaître que la chaîne de caractères "*NomAuteur*" est proche de "*Nom auteur*"

***Avantages :*** *indépendant des langues et des domaines.*

***Limite :***




- ***Solution :*** utiliser des informations contextuelles (description en langage naturel, noeuds et feuilles, etc.). Combinaison mesures lexicales et contextuelles.

# Avantages et limites des mesures fondées sur les chaînes de caractères

- ***Limite :*** problème de polysémie.
  Exemple, *souris ! (cf cours précédent)*
  - *Par exemple, polysémie sur les acronymes/sigles (cf prochain cours).*

    *Exemple :* JO = "Jeux Olympiques" *ou* "Journal Officiel"
    - Difficultés :
      - Mise à jour nécessaire des lexiques utilisés (acronymes récents).
      - Choix de l'acronyme adapté parmi une liste donnée.
      - Connaître la définition des acronymes des domaines spécialisés

# Reconnaissance de la langue et Recherche d'Information

- Autre utilisation des n-grammes pour les tâches de classification : **Reconnaissance de la langue** (étape préliminaire très efficace avant les approches de classification).

- Autre méthode de reconnaissance de la langue fondée sur l'identification des mots-outils propres aux langues (par exemple, “the”, “and”, etc.).

# Reconnaissance de la langue et Recherche d'Information

- *Exemple à partir d'un corpus parallèle :*

**`Adoption of the Minutes of the previous sitting`**

**`Adoption du procès-verbal de la séance précédente`**

***Exercice :***

- *Calculer les n-grammes sur ces deux phrases parallèles avec n = 2, 4, 10.*
- *Conclure sur l'utilisation des n-grammes de caractères pour la reconnaissance de la langue sur cet exemple.*
- *Généraliser ces résultats en donnant des exemples typiques de n-gramme du français et de l'anglais*