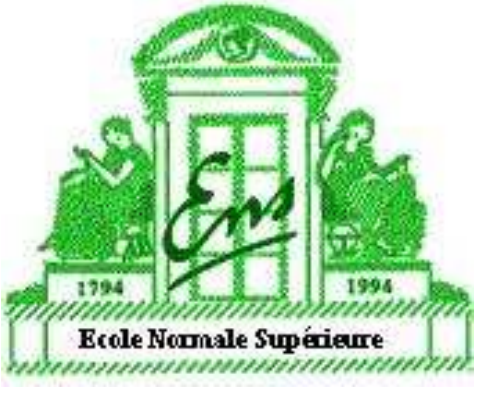
Découverte automatique de mots similaires

Exemple de techniques de fouille de textes

Pierre Senellart



Introduction : Problématique

- But ultime : recherche automatique de synonymes.
- Problème plus réaliste : on recherche des *mots similaires* ou des *quasi-synonymes*
- Applications :
  - Langage naturel
  - Recherche documentaire
  - Aide aux lexicographes

Plan de l'exposé

1. Introduction

2. Grands corpus
   - ✓ Principe
   - ✓ Espace vectoriel / documents
   - ✓ Thésaurus de mots rares
   - ✓ SEXTANT

3. Web
   - ✓ Particularités
   - ✓ Test de synonymie

4. Dictionnaires monolingues
   - ✓ Graphe du dictionnaire
   - ✓ Algorithmes considérés
   - ✓ Résultats expérimentaux

5. Conclusion

Principe

- Principe général : *Deux mots similaires sont deux mots qui apparaissent dans les mêmes contextes.*
- Définition de contexte ?
- Quelle mesure de similarité ?

Espace vectoriel orienté documents

- Dimensions : documents
  Vecteurs : termes

- Deux mesures de similarité envisagées :

$$cos(\mathbf{i},\mathbf{j}) = \frac{\mathbf{i}\cdot\mathbf{j}}{\sqrt{\mathbf{i}\cdot\mathbf{i}\times\mathbf{j}\cdot\mathbf{j}}}$$

$$cluster(\mathbf{i},\mathbf{j}) = \frac{\mathbf{i}\cdot\mathbf{j}}{\|\mathbf{i}\|_1}$$

Thésaurus de mots rares

- Dimensions : termes
  Vecteurs : documents

- Clustering des documents suivant une mesure de similarité.

- Recherche des discriminants indifférents à l'intérieur de chaque classe

- Approximation : les mots rares sont des discriminants indifférents

SEXTANT

Contextes : relations syntaxiques entres mots (un nom est le sujet d'un verbe, un nom est modifié par un adjectif, etc.)

$$poids(att) = 1 + \sum_{\text{nom } \mathbf{i}} \frac{p_{att,\mathbf{i}} log(p_{att,\mathbf{i}})}{log(\text{nb total de relations})}$$

$$p_{att,\mathbf{i}} = \frac{\text{nb de fois } att \text{ app. avec } \mathbf{i}}{\text{nb total d'attributs de } \mathbf{i}}$$

$$jac(\mathbf{i},\mathbf{j}) = \frac{\sum_{att \text{ attribut des deux } \mathbf{i} \text{ et } \mathbf{j}} poids(att)}{\sum_{att \text{ attribute de soit } \mathbf{i} \text{ soit } \mathbf{j}} poids(att)}$$

Particularités du Web

- Taille énorme, intraitable.
- Longs temps d'accès mais bonne indexation.
- Pourtant : corpus le plus vivant et le plus riche qui existe

Test de synonymie sur le Web

j est-il un bon synonyme de **i** ?

$$score_1(\mathbf{j}) = \frac{hits(\mathbf{i}\ AND\ \mathbf{j})}{hits(\mathbf{j})}$$

$$score_2(\mathbf{j}) = \frac{hits(\mathbf{i}\ NEAR\ \mathbf{j})}{hits(\mathbf{j})}$$

$$score_3(\mathbf{j}) = \frac{hits((\mathbf{i}\ NEAR\ \mathbf{j})\ AND\ NOT\ ((\mathbf{i}\ OR\ \mathbf{j})\ NEAR\ \text{“not”})}{hits(\mathbf{j}\ AND\ NOT\ (\mathbf{j}\ NEAR\ \text{“not”})}$$

$$score_4(\mathbf{j}) = \frac{hits((\mathbf{i}\ NEAR\ \mathbf{j})\ AND\ \mathbf{context}\ AND\ NOT\ ((\mathbf{i}\ OR\ \mathbf{j})\ NEAR\ \text{“not”})}{hits(\mathbf{j}\ AND\ \mathbf{context}\ AND\ NOT\ (\mathbf{j}\ NEAR\ \text{“not”})}$$

Graphe du dictionnaire

- Nœuds : mots définis dans le dictionnaire
- Arêtes : il y a une arête entre le mot a et le mot b si b apparaît dans la définition de a
- Graphe de voisinage : le graphe de voisinage d'un mot a est le sous-graphe formé de a, des fils de a et des parents de a.

Distance entre vecteurs

- A matrice d'adjacence du graphe.
- Distance entre le mot d'indice i et le mot d'indice j :

$$\|(A_{i,.} - A_{j,.})\| + \|(A_{.,i} - A_{.,j})^T\|$$

($\|\ \|$ est une norme vectorielle donnée).

- Proche des méthodes classiques de fouille de textes avec mesure de similarité.

ArcRank

- PageRank (Google) : distribution stationnaire des poids des nœuds du graphe correspondant au vecteur propre principal de la matrice d'adjacence.

- ArcRank :

$$r_{s,t} = \frac{p_s / |a_s|}{p_t}$$

$|a_s|$ est le degré sortant de s. p_t est le pagerank de t.

- *Les meilleurs synonymes de i sont les extrémités des meilleurs arcs arrivant sur ou quittant le noeud i.*

Comparaison de graphes

- On compare le graphe du dictionnaire avec le graphe $1 \longrightarrow 2 \longrightarrow 3$.

- Les bonds nœuds 2 sont les nœuds pointés par des bons nœuds 1 et pointant vers des bon nœuds 3 : définition mutuellement récursive

- Mots similaires : mots semblables à 2 dans le graphe du voisinage

Résultats expérimentaux - disappear

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	vanish	vanish	epidemic	vanish	vanish
2	wear	pass	disappearing	go away	cease to exist
3	die	die	port	end	fade away
4	sail	wear	dissipate	finish	die out
5	faint	faint	cease	terminate	go
6	light	fade	eat	cease	evaporate
7	port	sail	gradually		wane
8	absorb	light	instrumental		expire
9	appear	dissipate	darkness		withdraw
10	cease	cease	efface		pass away
Mark	3.6	6.3	1.2	7.5	8.6
Std dev.	1.8	1.7	1.2	1.4	1.3

Résultats expérimentaux - majesté

1. grandeur (0.382138)
2. titre (0.366114)
3. dignité (0.274143)
4. noblesse (0.267147)
5. noble (0.243879)
6. altesse (0.14966)
7. sire (0.147181)
8. gloire (0.134615)
9. pouvoir (0.125155)

1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires
5. Conclusion
   ✓ Résumé

En résumé…

1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires
5. Conclusion
   ✓ Perspectives

Perspectives