Modèle de langage en RI

1

Modèle de langage

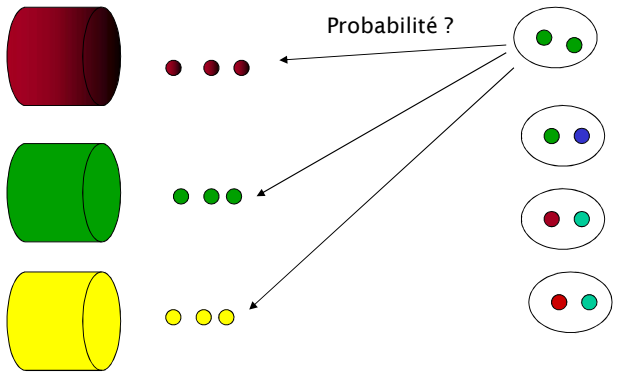
- Modèles de langues (ou de langages) tentent de modéliser « l'agencement de mots dans une langue »
- Exemple introductif
  - En Sibérie, il fait ….
  - In ……
- Probabilité de distribution de mots (une séquence de mots) dans un texte
  - Probabilité d'une séquence de mots dans une « langue » donnée.
  - p1=P (Dans ce cours nous allons étudier le modèle de langues)
- Notation
  - M la langue (le langage)
  - s une séquence
  - P(s/M) : probabilité d'observer s dans M

2

Métaphore

Source (génération de boules)

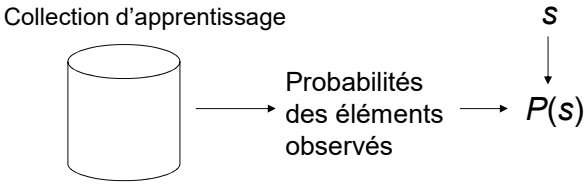
Quelle est la source qui a généré ces boules ?



3

Modèle de langage

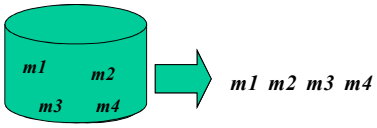
- Définition :
  - Un modèle de langage désigne une fonction de probabilité qui assigne une probabilité à une séquence de mots dans une langue.
  - Ce calcul se fait sur un corpus d'apprentissage



4

Probabilité d'une séquence

- D'une façon générale, la probabilité de générer une séquence donnée est mesurée comme suit

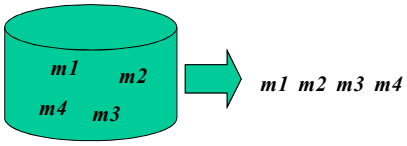


P (m1 m2 m3 m4) = P(m1) x P(m2/m1) x P(m3/m1 m2) x P(m4/m3 m2 m1)

5

Probabilité d'une séquence (suite)

- Si les événements sont indépendants



P (m1 m2 m3 m4) = P(m1) x P(m2) x P(m3) x P(m4)

6

Probabilité d'une séquence (suite)
Généralisation

- Modèle n-grammes (n séquence de mots ou caractères) $p(m_1, m_2, ..., m_l) = \prod_{i=1}^{l} P(m_i \mid m_1 ... m_{i-1})$
  - Uni-gramme $P(s) = \prod_{i=1}^{l} P(m_i)$
  - Bi-grammes $P(s) = \prod_{i=1}^{l} P(m_i \mid m_{i-1}) = \prod_{i=1}^{l} \frac{P(m_{i-1}, m_i)}{P(m_{i-1})}$
  - N-grammes $P(s) = \prod_{i=1}^{l} P(m_i \mid m_1 ... m_{i-1})$

7

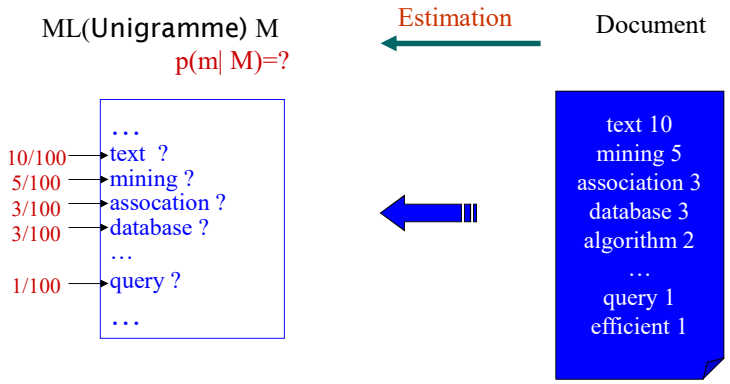
Estimation des probabilités

- Problème
  - Soit une séquence S, estimer son modèle de langage MS
- Modèle de langage de base
  - Maximum de vraisemblance

8

Maximum de vraisemblance (Maximum-likelihood)

- Compter la fréquence relative des mots m :
  - Pml(m|M) = #(m) / N



Un article sur le "text mining"
(total #mots=100)

9

Problème des fréquences Zéro

- Si un événement (un mot) n'apparait pas dans la séquence
  - Le modèle va assigner la probabilité 0 à l'événement

$$P(s) = \prod_{i=1}^{l} P(m_i) = 0, \quad si \quad \exists m_i / p(m_i / M) = 0$$

- Solution : assigner des probabilités différentes de zéro à de tels mots

10

Problème des fréquences Zéro (suite)

- Contraintes :
  - On ne peut pas assigner des valeurs différentes de zéro de manière aléatoire
  - La somme des probabilités de l'ensemble des événements doit être égale à 1.
  - Plusieurs solutions ….

Techniques de lissage

11

Techniques de lissage

- Méthodes de « discounting »
  - Laplace correction, Lindstone correction, absolute discounting, leavet one-out discounting, Good-Turing method
- Techniques d'Interpolation
  - Estimations de Jelinek-Mercer, Dirichlet

12

Méthodes de « discounting »

- Ajouter une constante (1, 0,5 ou ε) à toutes les fréquences
  - Laplace smoothing
    - Ajouter un à tous les événements (n-gram : s)

$$P_{add\_one}(s \mid M) = \frac{|s|+1}{\sum_{s_i \in V}(|s_i|+1)}$$

  - Lindstone Smoothing:
    - Ajouter ε puis normaliser

13

Méthodes de « discounting »

- **Dans MLE,** P(s|M) = #(s) / N

Avec : s = un n-gramme

#(s) = la fréquence de s. (problème avec #(s) = 0)

N= somme des fréquences des n-grammes

- **Good-Turing** : change la fréquence de s en

$$\#(s)^* = (\#(s)+1)\frac{n_{s+1}}{n_s}$$

Avec: #(s) = la fréquence de s

n_s = nombre de n-grammes de frequence #(s)

n_{s+1} = nombre de n-grammes de frequence #(s) *+1*

n_0= nombre total des n-grammes

Problème : tf* peut être zéro s'il n y a pas de n-grames de fréquence (#(s) *+1)*

14

- Soit s=« text mining information » et soit le document suivant

Calculer p(s\D) avec :

1/ MLE

2/Laplace smoothing (add_one)

3/ Good-Turing

15

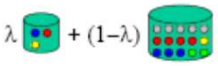
Lissage par interpolation

- Les méthodes de « discounting » traitent les mots qui n'apparaissent pas dans le corpus de la même manière. Or, il y a des mots qui peuvent être plus fréquents que d'autres
- Solution
  - Construire un modèle mixte : combiner deux modèles

16

Lissage par interpolation (suite)

- Interpolation (Jelinek-Mercer)
  - Combiner le modèle M , avec un modèle plus général



© james allan

  - Pb. "Règlage" de λ

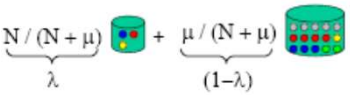
Dans le document *Dans le corpus*

$$P_{JM}(w_i \mid M) = \lambda . P_{ML}(w_i \mid M) + (1-\lambda) P_{JM}(w_i)$$

17

Lissage par interpolation (suite)

- Lissage de Dirichlet
  - Problème avec Jelinek-Mercer
    - Les documents long seront privilégiés
  - Prendre en compte la taille de l'échantillon
  - Si N est la taille de l'échantillon et μ une constante



© james allan

$$P_{Dir}(w_i \mid M) = \left(\frac{N}{N+\mu}\right) . P_{ML}(w_i \mid M) + \left(\frac{\mu}{N+\mu}\right) P_{Dir}(w_i)$$

18

Lissage par interpolation (suite)

- Lissage de Dirichlet (ex transp. 9)

$$P_{Dir}(w_i \mid d) = \frac{|d|}{|d| + \mu} \times \frac{tf(w_i, d)}{|d|} + \frac{|\mu|}{|d| + \mu} P_{ML}(w_i \mid C)$$

$$P_{Dir}(w_i \mid d) = \frac{tf(w_i, d) + \mu P_{ML}(w_i \mid C)}{|d| + \mu}$$

d: un document
C : le corpus (collection)
μ est une constante par exemple ½

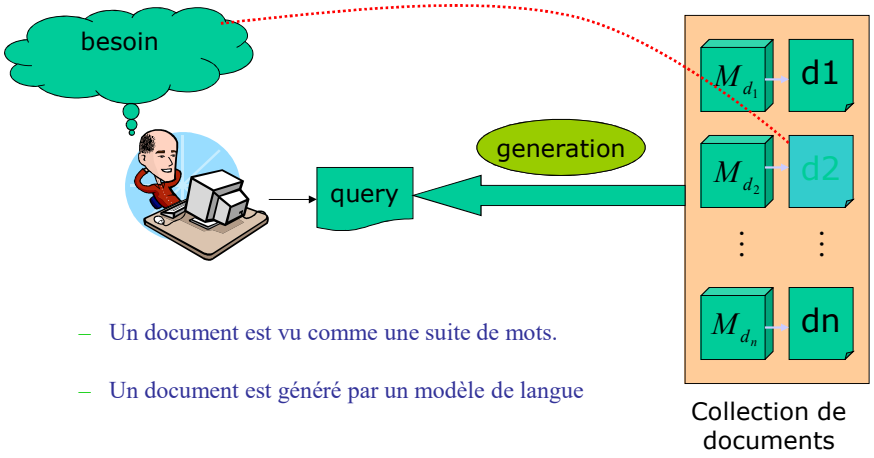
19

Modèle de Langage en RI

Plusieurs modèles, plusieurs adaptations

20

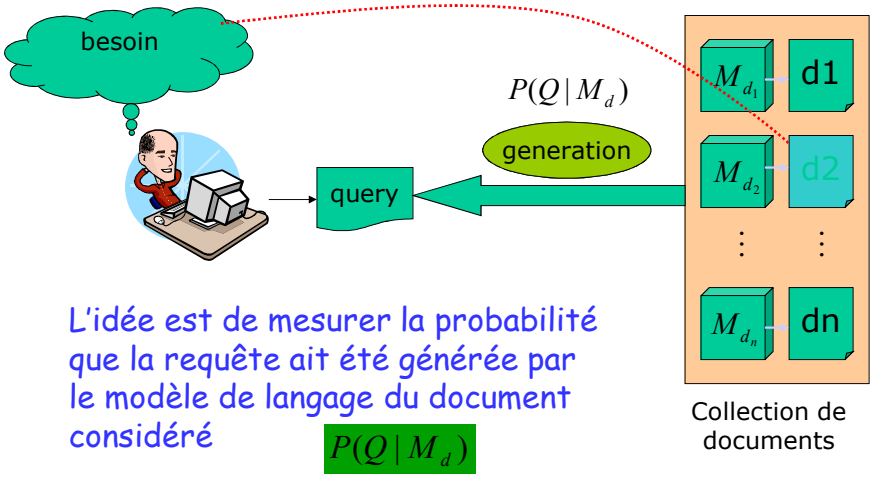
RI basée sur les ML



- Un document est vu comme une suite de mots.
- Un document est généré par un modèle de langue

21

….. et la RI là dedans



L'idée est de mesurer la probabilité que la requête ait été générée par le modèle de langage du document considéré $P(Q \mid M_d)$

22

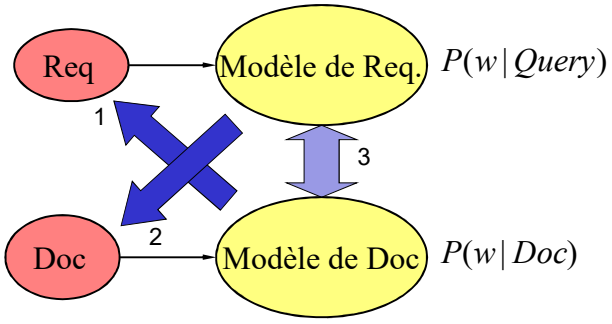
Comment estimer $P(Q/M_D)$

- Le modèle de langage est inconnu
  - Mais, nous avons un échantillon : le document,
  - Estimer le modèle de langage à partir de cet échantillon

23

ML en RI

- Il existe plusieurs manières d'adapter les ML à la RI.



3 Principes : Probabilité de générerer la requête à partir de Md (1),
Probabilité de générer le document à partir de Mq (2),
Combinaison (comparaison) des deux modèles(3)

24

ML en RI (suite)

- Principe 1: (principe standard) : génération de la requête par le document
  - RSV(D,Q)= $P(Q|M_D)$
  - Document D: représenté par son ML $P(w|M_D)$
  - Requête Q = séquence de mots $q_1,q_2,\ldots,q_n$
- Principe 2 : génération du document par la requête
  - RSV(D,Q)= P(D|Mq)
  - Requête Q: représentée par son ML P(w|Mq)
  - Document D = séquence de mots
- Principe 3: ratio de vraisemblance (comparaison de modèles)
  - Document D: LM $P(w|M_D)$
  - Requête Q: LM $P(w|M_Q)$
  - RSV(Q,D): comparaison entre $P(w|M_D)$ and $P(w|M_Q)$

25

Principe 1 : Génération de la requête par le document

- Chaque document est traité comme un modèle de langage
- Estimer le modèle de langage M_d de chaque document
- Classer les documents par leur probabilité de générer la requête

$$P(Q/M_D) = \prod_{t \in Q} P(t/M_D)$$

26

Principe 1 : génération de la requête par le document (suite)

- La probabilité de générer la requête sachant un modèle de langage du document d avec MLE est:

$$p(Q \mid M_d) = \prod_{t \in Q} p_{ml}(t \mid M_d)$$

$$= \prod_{t \in Q} \frac{tf_{(t,d)}}{dl_d}$$

Les termes de la requête sont générés de manière indépendantes

M_d : modèle de langage du document d
$tf_{(t,d)}$: fréquence d'un terme dans d
dl_d : nombre total de termes dans d

27

Approche1 : Problème avec l'Est. Max. Vrai. (MLE)

- Problème des tf = 0 : quand un document ne contient pas un ou plusieurs termes de la requête.
- Utiliser les techniques de lissage : le modèle mixte
  - Combiner le modèle de langage de document et le modèle de langage de la collection
  - Le modèle de collection est utilisé comme un modèle de référence pour les mots non observés dans le document (lissage JM)

$$p(Q,d) = \lambda p_{mle}(Q \mid M_d) + (1-\lambda) p(Q \mid M_c)$$

28

Approche1 : Modèle mixte de base

- Formulation générale

$$RSV(Q,d) = \prod_{t \in Q} ((1-\lambda)p(t/M_c) + \lambda p(t \mid M_d))$$

Modèle général (collection)

Modèle de document

$$p(t \mid M_c) = p(t) = \frac{total\_tf_t}{total\_tf\_col}$$

$total\_tf_t$: fréquence du terme dans la collection
total_tf : collection : nombre total de termes dans la collection

29

Exemple

- (2 documents)
  - d_1: Xerox reports a profit but revenue is down
  - d_2: Lucent narrows quarter loss but revenue decreases further
- Requête: *revenue down*
- MLE unigram;
  - Lissage JM $\lambda = ½$
  - Lissage Dir, $\mu = ½$

30

Principe 2 : Génération du document à partir de la requête

- Chaque requête est traitée comme un modèle de langage
- Estimer le modèle de langage M_q de chaque requête
- Classer les documents

$$P(D / M_Q) = \prod_{t \in Q} P(t / M_Q)$$

- Comme la requête est très réduite, ce type d'estimation n'est pas intéressant

31

références

J.M. Ponte and W.B. Croft. 1998. A language modelling approach to information retrieval. In *SIGIR 21*.

D. Hiemstra. 1998. A linguistically motivated probabilistic model of information retrieval. *ECDL 2*, pp. 569–584.

A. Berger and J. Lafferty. 1999. Information retrieval as statistical translation. *SIGIR 22*, pp. 222–229.

D.R.H. Miller, T. Leek, and R.M. Schwartz. 1999. A hidden Markov model information retrieval system. *SIGIR 22*, pp. 214–221.

M. Boughanem, W. Kraaij and J-Y. Nie. Modèles de langage pour la recherche d'information. 2004 (lavoisier)

32