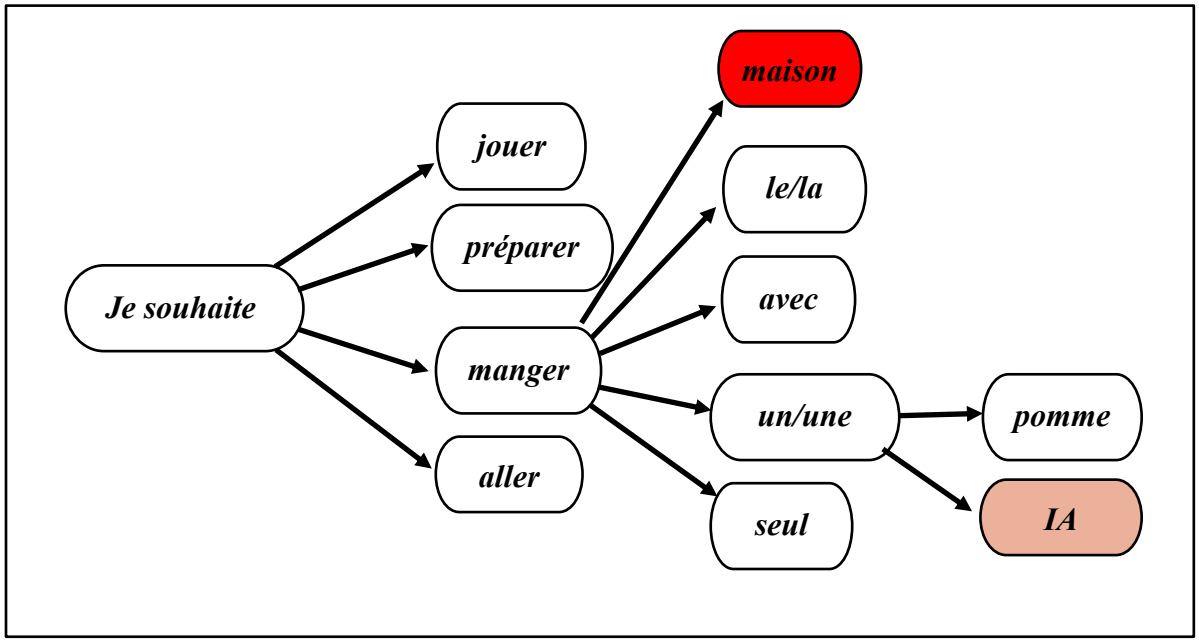
Chapitre 2 : Modèles statistiques de langage (Langue Model)

Ces sytsèmes de complétion ont un nom

- Modèle de langage - Modèle de langue – Language Model :

vise à prédire le mot suivant dans une séquence de mots avec une certaine probabilité.



Sachant un texte (une séquence de mots – contexte-) prédire le mot qui suivant

LM : (Language Model) - Modèle de langage
LLM : (Large Language Modele) - Grands Modèles de langage

Prise en compte de l'ordre des mots

- Un modèle de langage est un modèle probabiliste
  - Capable de prédire le mot manquant (suivant) dans une séquence de mots.
    - $P(w_{t+1}|w_1, w_2, \dots, w_t)$
    - Ex:
      - $P(x|en, sibérie, il, fait)$→ trouver x ?
  - Assigne la probabilité d'observer une séquence de mots dans une langue
    - $P(w_1, w_2, \dots, w_n)$ → probabilité d'observer la séquence $w_1, w_2, \dots, w_n$
    - Ex:
      - P (un garçon mange une pomme)
      - P (une pomme mange un garçon)

Comment estimer (calculer) ces probabilités ?

Prise en compte de l'ordre des mots

- Comment estimer (prédire) ces probabilités ?

Beaucoup de textes

Puis,

un simple comptage de séquences de mots (fréquences relatives de séquences)

Estimation du modèle

- Retour sur le calcul de probabilités
  - $P(w_1, w_2, \dots, w_m)$=P$(w_1) * P(w_2|w_1), \dots P(w_m|w_1, w_2, \dots, w_{m-1})$

$$P(w_1, w_2, \dots, w_m) = \prod_{i=1}^{m} P(w_i|w_1, w_2, \dots w_{i-1})$$

$$P(w_i|w_1, w_2, \dots w_{i-1}) = \frac{count(w_1, w_2, \dots w_{i-1}, w_i)}{count(w_1, w_2, \dots w_{i-1}, u\ )}$$

Estimation du modèle

- Modèles n-grammes → n étant la taille du modèle
  - soit s une observation (un texte) de *m* mots s=w_1 w_2…w_m
  - *Unigram* – (observe/génère des séquences de 1 mot)

$$P(w_1, w_2, \dots, w_m) = \prod_{i=1}^{m} P(w_i) = \frac{count(\, w_i \,)}{\left(\sum_{u \in V} [\![count(\, u)]\!]\right)}$$

  - *bigrams* – (observe/génère des séquences de deux mots)

$$P(w_1, w_2, \dots, w_m) = \prod_{i=1}^{m} P(w_i | w_{i-1}) = \frac{count(w_{i-1}\, w_i \,)}{\sum_{u \in V} count(w_{i-1} u)}$$

  - *Trigrams* – (génère/observe des séquences de 3 mots)

$$P(w_1, w_2, \dots, w_m) = \prod_{i=1}^{m} P(w_i | w_{i-2}, w_{i-1}) = \frac{count(w_{i-2} w_{i-1}\, w_i \,)}{\sum_{u \in V} count(w_{i-2} w_{i-1} u)}$$

Estimation du modèle (exemple)

- D'où viennent ces probabilités :
  - Modèle unigram

Text … …algorithm …
in a
Text
Ass
…

Text mining … is effecient …association … in database … to query. Text mining , … Text, .. Association mining …. …

Extraire les unigrams

Total des occurrences =30

text 10
mining 5
association 3
database 3
algorithm 2
query 1
efficient 1
……

Calcul de la probabilité d'une séquence

Estimation du Modèle

P(m)=

text ? 10/30
mining ? 5/30
association ? 3/30
database ? 3/30
algorithm? 2/30
query? 1/30
efficient? 1/30

Modèle de langage - Language Model

- D'où viennent ces probabilités :
  - Modèle bi gram ($P(w_i|w_{i\_1})$)

Text … …algorithm … in algorithm … to query. Text n… n…

Text mining is effecient association algorithm in database … to query. Text mining , … Text, .. Association rule …. …

P(mining/text), P(efficient/mining)

text mining 3 → count(text, mining) / count(text, *)
mining is 2 → count(mining, is) / count(mining, *)
association rule 1
is efficient 3
algorithm data 2
….
query process 1
……

Total des occurrences =20

Problème des fréquences nulles (zéro)

- Si un événement (un mot de la séquence) n'apparait pas dans le modèle, le modèle lui assigne une probabilité 0

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

- Solution : assigner des probabilités différentes de zéro aux événements (mots) absents
  - → Lissage (Smoothing)

Techniques de lissage

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

Lissage par interpolation

- Interpolation (Jelinek-Mercer)
  - Combiner le modèle M avec un modèle plus général (Modèle de référence)

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

  - Pb. “Règlage” de λ

Lissage par interpolation (JM)

- Modèle lissé (JM)

$$P_{JM}(t \mid d) = \alpha \times P_{ML}(t \mid d) + (1-\alpha) P_{ML}(t \mid C)$$

$$P(t \mid M_c) = p(t) = \frac{tf(t,C)}{\sum tf(t',C)}$$

Exemple

P("text retrieval")=P(text)*P(retrieval)=(10/30)*(0)

Collection 100 documents (total tf ds C (2000)
tf(retrieval,C)=6 → P_{ml}(retrieval/C)=6/2000
tf(text,C)=25→ P_{ml}(text|C)= 25/2000