Modélisations mathématiques

1. Construction et utilisation de modèles de langage probabilistes

Solen Quiniou

solen.quiniou@univ-nantes.fr

IUT de Nantes

Année 2020-2021 – Info 2

(dernière mise à jour : 8 décembre 2020)

Plan du cours

Plan du cours

Introduction de cette partie du module

- À quoi sert le traitement automatique des langues (TAL) ?
  - On dispose de plus en plus de masses de documents multimédia.
    - ⋆ Documents écrits, manuscrits, audio, vidéo...
  - Il faut des méthodes pour exploiter ces documents d'où le TAL.
    - ⋆ Reconnaissance de la parole, reconnaissance d'écriture manuscrite
    - ⋆ Recherche d'information, résumé automatique, traduction automatique, correction orthographique, catégorisation de textes...
- Pourquoi utiliser des modèles de langage probabilistes (ML) ?
  - Un modèle de langage probabiliste permet de donner une *probabilité* à une phrase pour dire si elle peut exister dans une langue donnée.
    - → Exemple : utilisation de ML pour trier des phrases par probabilités décroissantes.
  - Correction orthographique :
    - ⋆ P(nous venons à l'IUT) > P(nous venont à l'IUT)
  - Traduction automatique :
    - ⋆ P(quel âge as-tu ?) > P(combien vieux es-tu ?)
  - Reconnaissance de la parole :
    - ⋆ P(j'aime les maths) > P(James lait mat)

Plan du cours

Modèles de langage probabilistes

En pratique, ne tient pas la route

- Le but d'un modèle de langage probabiliste est de calculer la probabilité d'une phrase s, représentée par une séquence de mots $w_1 \dots w_i \dots w_N$:

$$P(s) = P(w_1 \dots w_N) = P(w_1, \dots, w_N)$$

et dans cet ordre là

- En utilisant la règle des probabilités en chaîne, cette probabilité jointe peut être réécrite en un produit de probabilités conditionnelles :

$$P(s) = P(w_1) \times P(w_2|w_1) \times \dots \times P(w_N|w_1 \dots w_{N-1}) = \prod_{i=1}^{N} P(w_i|w_1 \dots w_{i-1})$$

  - $h_i = w_1 \dots w_{i-1}$ est appelé historique du mot w_i.
  - Les probabilités $P(w_i|h_i)$ sont les paramètres du modèle de langage.
    $\rightarrow$ Elles sont estimées sur des corpus (grandes collections de textes).

- Remarque
  - En pratique, nous n'utiliserons pas des probabilités mais des log-probabilités (logarithme décimal des probabilités) pour éviter les débordements en multipliant des probabilités très faibles et pour passer à des additions :

(sur des fichiers de plusieurs milliards de mots, les probabilités sont trop petites)

$$P_1 \times P_2 \times P_3 \times P_4 \propto \log P_1 + \log P_2 + \log P_3 + \log P_4$$

Exemple de modèle de langage probabiliste

Dans le corpus suivant, les étiquettes *<s>* et *</s>* servent à identifier les débuts et fins de phrase, respectivement [1].

nouveau vocabulaire : "mot début de phrase", "mot fin de phrase"

Corpus

en général on ne garde pas la casse

1. *<s> Antoine écoute Thom </s>*
2. *<s> Denis écoute une autre chanson </s>*
3. *<s> Elle écoute une chanson de Lionel </s>*

1. La probabilité de la phrase « *<s> Antoine écoute Thom </s>* » se calcule ainsi :
   $P(\text{<s> Antoine écoute Thom </s>}) = P(\text{<s>}) \times P(\text{Antoine}|\text{<s>}) \times P(\text{écoute}|\text{<s> Antoine}) \times P(\text{Thom}|\text{<s> Antoine écoute}) \times P(\text{</s>}|\text{<s> Antoine écoute Thom})$

2. Pour calculer la probabilité de la phrase « *<s> Lionel écoute une chanson </s>* », les probabilités suivantes doivent avoir été estimées, dans le modèle de langage : $P(\text{<s>})$, $P(\text{Lionel}|\text{<s>})$, $P(\text{écoute}|\text{<s> Lionel})$, $P(\text{une}|\text{<s> Lionel écoute})$, $P(\text{chanson}|\text{<s> Lionel écoute une })$, $P(\text{</s>}|\text{<s> Lionel écoute une chanson})$

→ Seule la probabilité $P(\text{<s>})$ peut réellement être estimée sur ce corpus !

1. *<s>* et *</s>* sont considérés comme des mots supplémentaires du vocabulaire.

Limitations des modèles de langage probabilistes

Exemple de corpus

1. *<s> Antoine écoute Thom </s>*
2. *<s> Denis écoute une autre chanson </s>*
3. *<s> Elle écoute une chanson de Lionel </s>*

- Problème du nombre de probabilités à estimer
  - Dans ce corpus constitué de 3 phrases, il y a déjà 18 probabilités à estimer, ce qui est déjà beaucoup par rapport au nombre de mots et de phrases !
    - ⋆ 5 probabilités pour la première phrase
    - ⋆ 6 autres probabilités pour la deuxième phrase
    - ⋆ 7 autres probabilités pour la troisième phrase

→ Pour diminuer le nombre de probabilités à estimer, nous définissons des classes d'équivalence sur les historiques des mots.

- Problème de la généralisation et du manque de données
  - À partir des probabilités estimées sur le corpus, la phrase « *Antoine écoute une chanson* » a une probabilité nulle alors qu'elle est correcte en français :

    $P(\text{<s> Antoine écoute une chanson </s>}) =$
    $P(\text{<s>}) \times P(\text{Antoine}|\text{<s>}) \times P(\text{écoute}|\text{<s> Antoine}) \times P(\text{une}|\text{<s> Antoine écoute}) \times$
    $P(\text{chanson}|\text{<s> Antoine écoute une }) \times P(\text{</s>}|\text{<s> Antoine écoute une chanson}) = 0$

→ Pour éviter les probabilités nulles, nous effectuons un lissage des probas.

Plan du cours

Modèles de langage *n*-grammes

- Dans un modèle *n*-gramme, l'historique h_i d'un mot w_i est approximé par les $n-1$ mots qui le précèdent (d'après l'hypothèse de Markov).
  - L'historique devient $h_i = w_{i-n+1} \dots w_{i-1}$: deux historiques se terminant par les mêmes $n-1$ mots appartiennent alors à la même classe d'équivalence.
  - Une séquence de n mots est appelée un *n*-gramme.

  $\rightarrow$ n correspond à l'ordre du modèle *n*-gramme.

- La probabilité d'une phrase $s = w_1 \dots w_s$ se réécrit alors en

$$P(s) = \prod_{i=1}^{N} P(w_i|h_i) = \prod_{i=1}^{N} P(w_i|w_{i-n+1}^{i-1})$$

- Valeurs courantes de n
  - Modèle unigramme ($n = 1$)

    $P(\text{<s> il fait beau </s>}) = P(\text{<s>}) \times P(il) \times P(fait) \times P(beau) \times P(\text{</s>})$
  - Modèle bigramme ($n = 2$)

    $P(\text{<s> il fait beau </s>}) = P(\text{<s>}) \times P(il|\text{<s>}) \times P(fait|il) \times P(beau|fait) \times P(\text{</s>}|beau)$
  - Modèle trigramme ($n = 3$)

    $P(\text{<s> il fait beau </s>}) =$

    $P(\text{<s>}) \times P(il|\text{<s>}) \times P(fait|\text{<s> } il) \times P(beau|il\ fait) \times P(\text{</s>}|fait\ beau)$

Estimation des paramètres des modèles *n*-grammes

- L'estimation des probabilités $P(w_i|w_{i-n+1}^{i-1})$ se fait au maximum de vraisemblance (*Maximum Likelihood*), c'est-à-dire en comptant les fréquences relatives des *n*-grammes sur un corpus d'apprentissage :

$$P(w_i|w_{i-n+1}^{i-1}) = \begin{cases} 1 & \text{si } w_i = <s> \\ \frac{c(w_{i-n+1}^{i-1} w_i)}{\sum_{w_j \in V} c(w_{i-n+1}^{i-1} w_j)} & \text{si } \sum_{w_j \in V} c(w_{i-n+1}^{i-1} w_j) \neq 0 \\ 0 & \text{sinon} \end{cases}$$

  - n : ordre du modèle *n*-gramme ;
  - V : vocabulaire considéré (c'est-à-dire ensemble des mots tous différents, incluant <s> et </s>) ;
  - $c(.)$: nombre d'occurrences du *n*-gramme en paramètre (dans le corpus d'apprentissage).

- Exemple : modèle bigramme sur le corpus précédent

$$P(Antoine|<s>) = \frac{c(<s> Antoine)}{\sum_{w_j \in V} c(<s> w_j)} = \frac{1}{3}$$

Remarques sur l'estimation des paramètres

- L'estimation des probabilités $P(w_i|w_{i-n+1}^{i-1})$ peut se réécrire ainsi :

$$P(w_i|w_{i-n+1}^{i-1}) = \begin{cases} 1 & \text{si } w_i = <s> \\ \frac{c(w_i)}{\sum_{w_j \in V} c(w_j)} & \text{si } n = 1 \text{ et } w_i \neq <s> \\ \frac{c(w_{i-n+1}^{i-1} w_i)}{c(w_{i-n+1}^{i-1})} & \text{si } n \geq 2, w_i \neq <s> \text{ et } c(w_{i-n+1}^{i-1}) \neq 0 \\ 0 & \text{sinon} \end{cases}$$

nombre d'occurence de ce mot après l'historique / nombre d'occurences totales apres l histotrique

→ Sur le corpus précédent, on a alors : $P(Antoine|<s>) = \frac{c(<s>\ Antoine)}{c(<s>)} = \frac{1}{3}$.

- Pour un historique $h_i = w_{i-n+1}^{i-1}$ donné, la somme des probabilités des n-grammes commençant par cet historique doit être égale à 1 :

$$\sum_{w_j \in V} P(w_j|w_{i-n+1}^{i-1}) = 1.$$

- En général, on ne prend pas en compte la casse des mots.
  - On commence toujours par mettre tous les mots en minuscule.

Exercice sur les modèles *n*-grammes

Corpus

1. *<s> Antoine écoute Thom </s>*
2. *<s> Denis écoute une autre chanson </s>*
3. *<s> Elle écoute une chanson de Lionel </s>*

1. Comment se décompose la probabilité de la phrase « *<s> Antoine écoute une chanson </s>* », dans un modèle unigramme ? Dans un modèle bigramme ? Dans un modèle trigramme ?

2. Estimez les paramètres du modèle unigramme. Combien y en a-t-il ?
3. Estimez les paramètres du modèle bigramme. Combien y en a-t-il ?
   - Commencez par identifier les probabilités nulles
   - Vérifiez enfin que, pour chaque historique, la somme des probabilités des *n*-grammes commençant par cet historique est égale à 1

4. Quelle est la probabilité de la phrase « *<s> Antoine écoute une chanson </s>* », par un ML unigramme ? Par un ML bigramme ?
5. Quelle est la probabilité de la phrase « *<s> Lionel écoute une chanson </s>* », par un ML unigramme ? Par un ML bigramme ?

Plan du cours

Lissage des probabilités

- Malgré la grande taille des corpus d'apprentissage, il est difficile d'y trouver toutes les séquences de mots valides dans la langue considérée.
  - Or, si un n-gramme est absent du corpus d'apprentissage, sa probabilité $P(w_i|w_{i-n+1}^{i-1})$ sera nulle et les phrases composées de ce n-gramme auront également une probabilité nulle.

- Les techniques de lissage permettent de n'avoir aucune probabilité $P(w_i|w_{i-n+1}^{i-1})$ nulle dans le ML. Elles procèdent en deux étapes :
  1. On réduit les probabilités des n-grammes présents dans le corpus d'apprentissage.
     - ⋆ Réduction absolue : même réduction pour tous les n-grammes.
     - ⋆ Réduction relative : la réduction d'un n-gramme dépend de sa fréquence.
  2. On redistribue sur l'ensemble des n-grammes du vocabulaire.
     - ⋆ Repli (*backoff*) : on utilise les probabilités d'un ML d'ordre inférieur, pour les n-grammes non observés.
     - ⋆ Interpolation : on combine les probabilités du ML avec celles des ML d'ordre inférieur (ainsi, pour un ML trigramme, on combine les probabilités du ML trigramme avec celles du ML bigramme et celles du ML unigramme).

→ Il existe de nombreuses techniques de lissage combinant différentes méthodes de réduction et de redistribution des probabilités.

Méthode de lissage de Laplace

- La méthode de lissage de Laplace (aussi appelée *Add-One Estimation*) repose sur la simple hypothèse que chaque *n*-gramme apparaît en réalité une fois de plus que la fréquence observée sur le corpus d'apprentissage.

- Les probabilités des *n*-grammes sont alors estimées par :

$$P_{Laplace}(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i-1} w_i) + 1}{\sum_{w_j \in V} c(w_{i-n+1}^{i-1} w_j) + |V| - 1}$$

V : taille du vocabulaire

  - Remarque : au dénominateur, on ajoute $|V| - 1$ car on ne compte pas <s> dans les mots w_j car <s> ne peut pas voir d'autres mots dans son historique

- La probabilité d'une phrase *s* reste inchangée et s'écrit ainsi :

$$P(s) = \prod_{i=1}^{N} P_{Laplace}(w_i|w_{i-n+1}^{i-1})$$

Exercice sur le lissage de Laplace

Corpus

1. *<s> Antoine écoute Thom </s>*
2. *<s> Denis écoute une autre chanson </s>*
3. *<s> Elle écoute une chanson de Lionel </s>*

1. Estimez les paramètres du modèle unigramme avec lissage de Laplace.
2. Estimez les paramètres du modèle bigramme avec lissage de Laplace et vérifiez que, pour chaque historique, la somme des probas des n-grammes commençant par cet historique est égale à 1.

3. Quelle est la probabilité de la phrase « *<s> Antoine écoute une chanson </s>* », par un ML unigramme avec lissage de Laplace ? Par un ML bigramme avec lissage de Laplace ?
4. Quelle est la probabilité de la phrase « *<s> Lionel écoute une chanson </s>* », par un ML unigramme avec lissage de Laplace ? Par un ML bigramme avec lissage de Laplace ?

Plan du cours

Évaluation des modèles de langage

- L'évaluation intrinsèque d'un modèle de langage est obtenue en calculant sa perplexité sur un corpus de test T :

$$PPL_{ML}(T) = 2^{H_{ML}(T)},$$

  - $H_{ML}(T)$ est l'entropie croisée du ML sur le corpus T :

$$H_{ML}(T) = -\frac{1}{|T|}\log_2(P(T)).$$

- La perplexité mesure le pouvoir prédictif du modèle de langage : $PPL_{ML}(T) = k$ signifie que le ML hésite en moyenne entre k mots.

$\rightarrow$ En résumé, plus la perplexité est faible et meilleur est le ML.

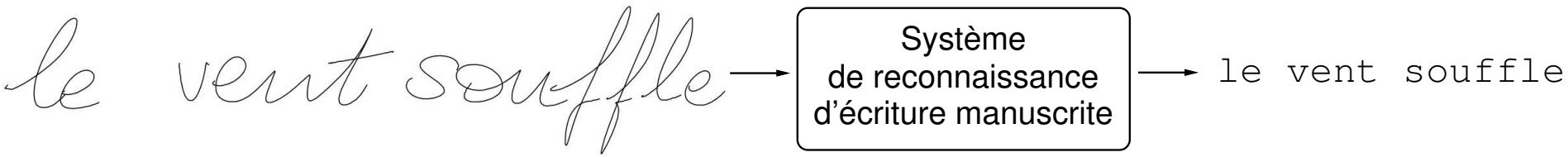
- L'évaluation extrinsèque d'un modèle de langage s'effectue par rapport aux performances de l'application utilisant le ML.
  - Par exemple, pour comparer deux ML utilisés pour faire de la traduction, on compare le nombre de mots correctement traduits avec chaque ML.

Plan du cours

Système de reconnaissance d'écriture manuscrite

- La reconnaissance de l'écriture manuscrite est un exemple de l'utilisation de modèles de langage.

- Un système de reconnaissance d'écriture manuscrite permet de transformer un texte manuscrit (représenté soit par une image, soit par une succession de points) en le texte qui a été écrit par l'utilisateur.



Création d'un système de reconnaissance de phrases

1. Construction du système de reconnaissance
   - Apprentissage des modèles de langage sur des corpus d'apprentissage.
   - Écriture d'un algorithme de reconnaissance de phrase qui utilise les modèles de langage et apprentissage de paramètres sur un ensemble de phrases d'apprentissage.

2. Évaluation du système de reconnaissance
   - Calcul des performances du système de reconnaissance, en comparant la phrase produite par le système (appelé hypothèse) et la phrase attendue (appelé référence), sur un ensemble de phrases de test.

Utilisation du système de reconnaissance

- Une fois le système de reconnaissance final choisi parmi les différents systèmes construits (après évaluation des performances de chacun d'eux), l'utilisation du système de reconnaissance se fait simplement : il suffit de donner la phrase manuscrite en entrée du système et le système de reconnaissance produira la phrase résultat, en utilisant l'algorithme de reconnaissance et les modèles de langage.

- Remarques sur la construction des systèmes de reconnaissance
  - On construit tout d'abord un premier système qui utilise un algorithme simple, pour la reconnaissance.
  - On construit ensuite un deuxième système qui utilise un algorithme plus complexe, notamment pour pouvoir reconnaître les phrases qui n'avaient pas été reconnues par le premier système.
  - ...