Modélisation statistique de la langue

Reconnaissance automatique de la parole, Traduction automatique

Alexandre Allauzen
allauzen@limsi.fr

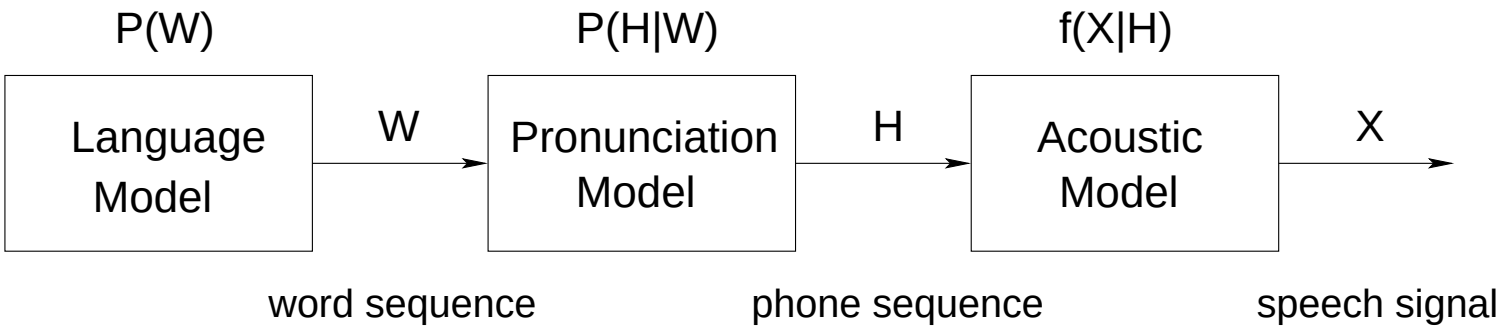
Université Paris Sud / LIMSI-CNRS

Pour aujourd'hui

1. Introduction
2. Grammaire formelle
3. Modèle n-gram
4. Estimation robuste (smoothing)

Plan

Approche statistique en reconnaissance automatique de la parole



L'équation fondamentale

En reconnaissance automatique de la parole, l'objectif est de déterminer la séquence de mots qui maximise la probabilité *a posteriori* :

$$\begin{aligned} \hat{W} &= \operatorname*{argmax}_{W} P(W|X) = \operatorname*{argmax}_{W} P(X|W)P(W) \\ &= \operatorname*{argmax}_{W} P(W) \sum_H P(H|W) f(X|H) \end{aligned} \tag{1}$$

Modélisation générative

L'équation fondamentale

$$\hat{W} = \underset{W}{\text{argmax}}\, P(W) \sum_{H} P(H|W) f(X|H)$$

Vision générative, source/canal, *noisy channel*

- $\boldsymbol{W}$ est généré par un modèle linguistique $P(W)$

  Modèles de langage

- Le modèle de prononciation $P(H|W)$ le transforme en une séquence de phonèmes H

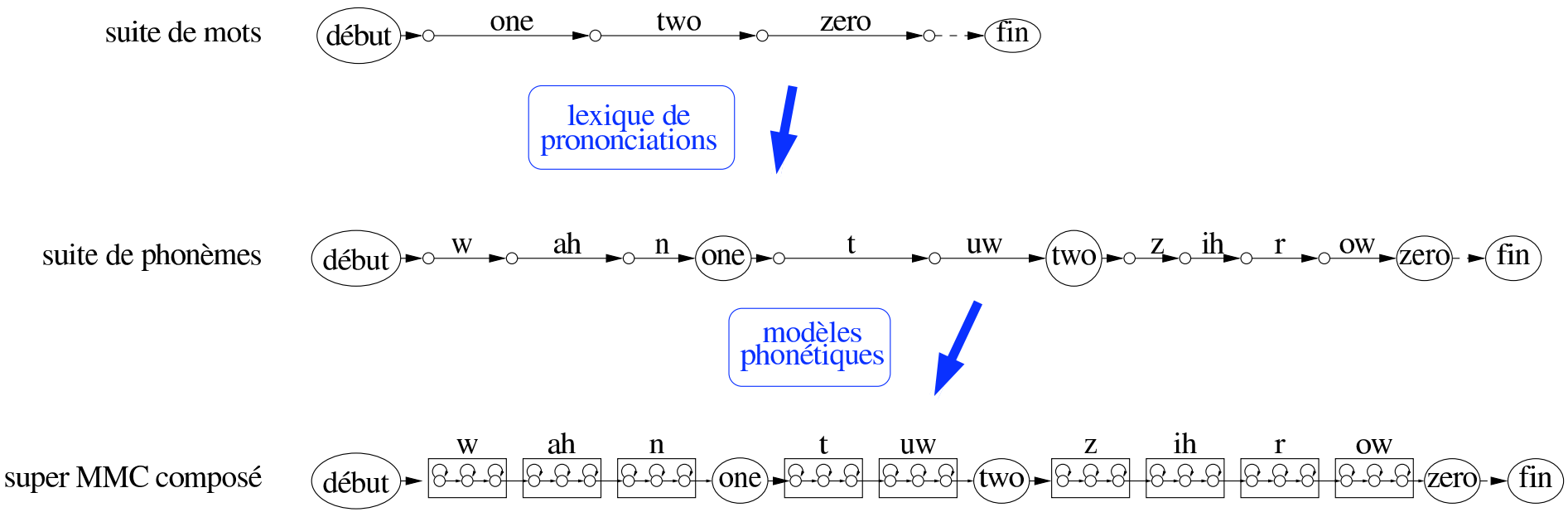
  Lexique de prononciation

- Encodée par le canal acoustique $\boldsymbol{f(X|H)}$ dans le signal $\boldsymbol{X}$

  Modèles acoustiques

- Le décodage : *argmax*

Modélisation générative



Un modèle de langage, est-ce utile ?

Prenons un exemple : ”Tu vois ce convoi ? ”

Pourquoi ne pas mettre ces mots sur le signal :

tu	vois	ce	qu'on	voit ?
tue	voie	ce	qu'on	voit ?
tues	vois	se	qu'on	voit ?
tu	vois	se	qu'on	voix ?
tu	vois	ceux	qu'on	voit ?
tu	vois	ceux	convoi ?	

Les Connaissances nécessaire sur le langage

- Morpho-syntaxique,
- Sémantique et pragmatique (*le chat boit son thé*),
- Le contexte, les thématiques (*convoi* ?).

Complexité du langage

Difficultés inhérentes aux langages apparaissent à différents niveaux :

graphémique

absence de voyelles en arabe écrit

lexical

mots composés en allemand ou encore majuscules au début de tous les substantifs

morpho-syntaxique

- nombre élevé de flexions en français
- un verbe finlandais peut connaître plus de 10 000 formes
- le pluriel malaisien est la répétition

et la sémantique ?

Il y a peu d'eau dans les os.

Alors modélisons !

Historiquement, deux approches s'affrontaient :

Noam Chomsky, 1969

It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Alors modélisons !

Historiquement, deux approches s'affrontaient :

Noam Chomsky, 1969

It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Frederick Jelinek, 1988

Whenever I fire a linguist our system performance improves.

Alors modélisons !

Historiquement, deux approches s'affrontaient :

Noam Chomsky, 1969

It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Frederick Jelinek, 1988

Whenever I fire a linguist our system performance improves.

puis en 2004

Some of my best friends are linguists.

Alors modélisons !

Historiquement, deux approches s'affrontaient :

Noam Chomsky, 1969

It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Frederick Jelinek, 1988

Whenever I fire a linguist our system performance improves.

puis en 2004

Some of my best friends are linguists.

Dit autrement

- Inférer les connaissances du langage humain sur les données.
- Partir des données - d'observations - pour faire émerger via des modèles statistiques des connaissances "non-supervisée" ou "semi-supervisée".

Modèles linguistiques, modèles de langage

Grammaire formelle *vs* modèle probabiliste

Grammaires formelles :

règles définissant l'ensemble des constructions linguistiques possibles.

Avantage

suffisantes pour les langages artificielles (langage de programmation) ou suffisamment contraints (ex. les nombres, dates...).

Inconvénients

difficiles à mettre en œuvre pour le langage naturel

- coûteuses (règles expertes)
- problèmes de couverture (sur et sous-génération)
- phrases agrammaticales non-admises

Modèles linguistiques...

Grammaires probabilistes :

modéliser les régularités statistiques dues aux contraintes lexicales, syntaxiques et sémantiques.

Estimation de probabilités d'émission

$P(\boldsymbol{W})$ ou $P(\boldsymbol{S}|\boldsymbol{W})$, avec $\boldsymbol{W}$ une séquence de mots, $\boldsymbol{S}$ la structure cachée associée.

Le modèle *n*-gram

association de probabilités aux suites de mots ou de catégories grammaticales

- permettent de traiter des phrases agrammaticales,
- simples à mettre en œuvre,
- apprentissage automatique,
- besoin de corpus importants.

Motivations

- Reconnaissance automatique de la Parole
- Reconnaissance manuscrite en ligne
- Reconnaissance de caractères
- Corrections orthographiques, re-accentuation de textes
- Génération de textes, traduction, aide à l'apprentissage des langues
- Dialogue, segmentation thématique, recherche documentaire
- Fouille de données textuelles, audiovisuelles

Plan

Plan

Génération *vs* reconnaissance

Mécanisme de génération

Nous sommes capables d'énoncer des phrases correctes sans les avoir jamais observées auparavant.

Mécanisme de reconnaissance

Nous sommes capables de reconnaître des phrases correctes sans les avoir jamais observées auparavant.

→ Tri automatique des phrases correctes/autres séquences de mots

Comment décrire ce mécanisme?
Il faudrait établir une **théorie exhaustive** du français

Analyseurs *vs* grammaires génératives

Analyseur de langage

Le but est de déterminer, au moyen d'un algorithme déterministe (donc se terminant toujours au bout d un temps fini), si une phrase donnée appartient au langage.

- p.ex: compilateurs (partie analyse lexicale + syntaxique)
- résultats: statut (succès/échec) + arbre syntaxique

Générateur de langage (formel)

Ensemble de règles pour générer toutes les phrases valides possibles du langage

- "grammaires génératives"
- souvent plus facile à comprendre pour les humains

Grammaires formelles

Pour les langues naturelles :
$\nexists$ formalisation complète du mécanisme de génération et de reconnaissance
$\exists$ approximations :
« *Phrase Structure Grammar* » (**N. Chomsky**, 1959)

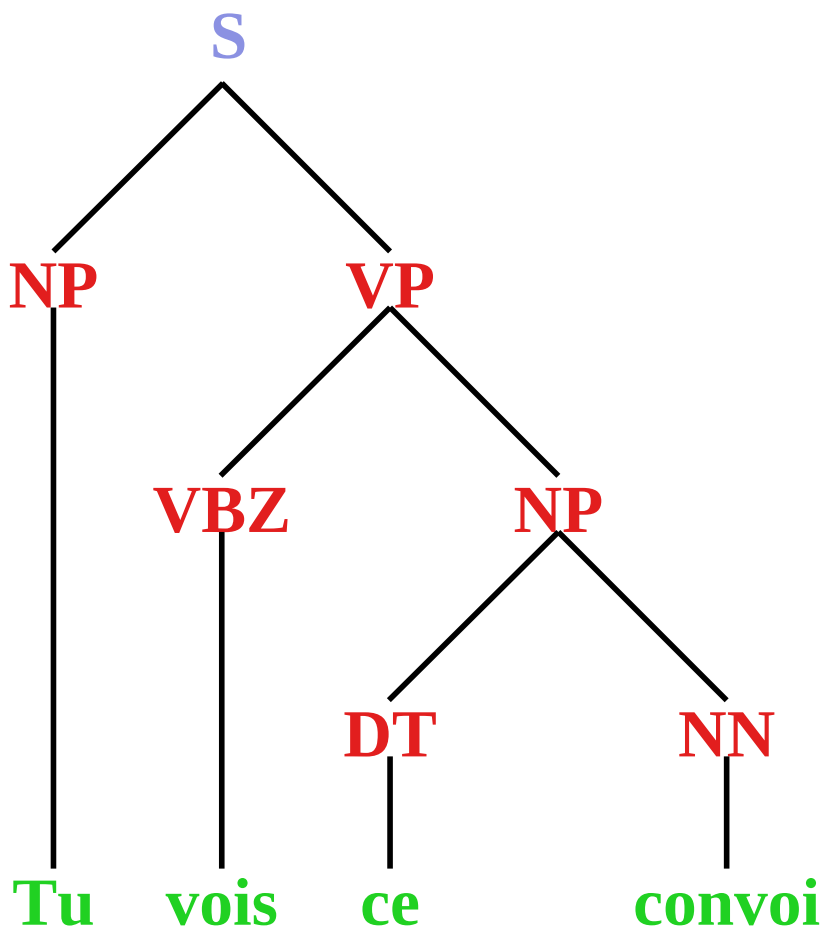
D'après Chomsky, le cerveau humain possède une faculté innée pour le langage. Cette faculté correspondrait à une grammaire universelle, c-à-d un ensemble de principes communs à tous les langages humains. Les travaux de Chomsky visent à décrire cette grammaire universelle.

Définition d'une grammaire formelle

Une grammaire formelle c'est un quadruplet.

- **N** ensemble de symboles non-terminaux
- **T** ensemble de symboles terminaux
- **R** ensemble de règles d'écriture
- **S** le symbole de départ

$$G = (N , T , R , S)$$



Classification de Chomsky

- La définition des grammaires génératives donnée ci-dessus n'impose aucune contrainte sur les productions.
- En introduisant des limitations sur la forme de ces productions, Noam Chomsky a introduit en 1956 une classification hiérarchique des grammaires et des langages, très généralement acceptée (de 0 à 3).
- Chomsky s'intéresse avant tout aux langues naturelles, mais il n'en constitue pas moins un pionnier de l'informatique !

Plan

Exemple de grammaire formelle

Le langage des expressions arithmétiques

- N = {expr, nombres, op,chiffres}
- T = {0,1,2,3,4,5,6,7,8,9,+,*,-,/}
- R ={ expr $\rightarrow$ expr op expr,
  op $\rightarrow$ +, op $\rightarrow$ -, op $\rightarrow$ *, op $\rightarrow$ /,
  expr $\rightarrow$ nombres,
  nombres $\rightarrow$ nombres chiffres,
  chiffres $\rightarrow$ 0, chiffres $\rightarrow$ 1, chiffres $\rightarrow$ 2, chiffres $\rightarrow$ 3, chiffres $\rightarrow$ 4,
  chiffres $\rightarrow$ 5, chiffres $\rightarrow$ 6, chiffres $\rightarrow$ 7, chiffres $\rightarrow$ 8, chiffres $\rightarrow$ 9 }

Que dire de : 1 + 2 / 3 ??

Systèmes générateurs

Grammaires : systèmes formels générateurs
Les expressions bien formées (ou phrases) d'une langue $\mathcal{L}$ sont obtenues (ou engendrées) à partir d'un symbole initial, en appliquant un ensemble de productions (ou règles de formation)

```
S       → Sujet Verbe
Sujet   → Article Nom
Article → l' | le
Nom     → étudiant | enseignant | chercheur
Verbe   → étudie | avance | travailles | écoute | enseigne
```

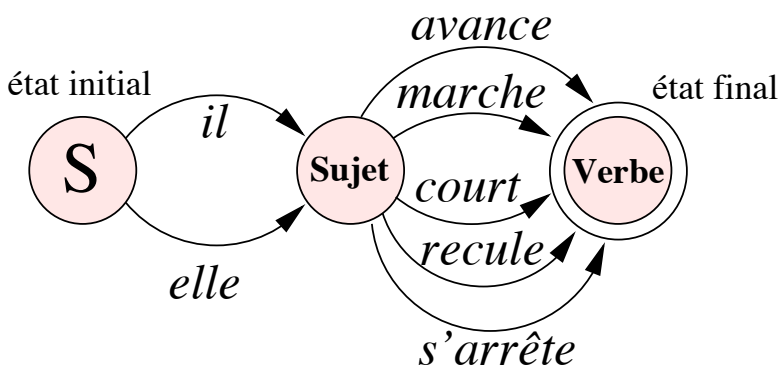
Exemples de phrases admissibles dans $\mathcal{L}$: *l'étudiant étudie, l'enseignant enseigne, . . .* mais aussi *le étudiant enseigne, l' enseignant travailles, . . .*

Quid de la sémantique ?

Systèmes accepteurs

Automates : systèmes formels accepteurs

En partant d'une phrase, leurs règles vont vérifier si cette phrase est ou non valide par rapport au langage donné



il marche $\in \mathcal{L}$
elle baille $\notin \mathcal{L}$

Automate $\equiv$ grammaire

Plan

Plan

Notation, mise en équation

Encore une fois

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} \mathrm{P}(\boldsymbol{W}/\boldsymbol{X}) = \underset{\boldsymbol{W}}{\operatorname{argmax}} \sum_{\boldsymbol{H}} \mathrm{P}(\boldsymbol{W})\mathrm{P}(\boldsymbol{H}|\boldsymbol{W})\boldsymbol{f}(\boldsymbol{X}|\boldsymbol{H}) \quad (2)$$

Notation

- W est une séquence de variable aléatoire (V.A) : $W = \{W_1, W_2, \ldots, W_N\}$.
- Chaque V.A est construite sur le même espace de réalisation : le vocabulaire V.
- La réalisation d'une V.A se note $W_i = w_i$.
- Abus d'écriture, omission de la référence à la V.A.

Définitions

Modèle de langage

Le modèle de langage assigne une probabilité non nulle à toutes séquences de mots $\boldsymbol{W}$ extraites du vocabulaire $\boldsymbol{V}$

Définition : le vocabulaire et le mot

$\boldsymbol{V}$ = liste des mots qui peuvent être reconnus par le système + $< UNK >$.
Un mot est une suite finie et ordonnée de caractères

$$\boldsymbol{W} = (w_1, w_2, ..., w_n) \text{ , avec } \quad w_i \in \boldsymbol{V}$$
$$P(\boldsymbol{W}) = \prod_{i=1}^{T} P(w_i | w_1, w_2, ... w_{i-1}) \tag{3}$$

Corpus d'apprentissage ou d'entraînement

Estimation des probabilités à partir d'observation sur des corpus d'entraînement (pour toutes les séquences ?)

Classe d'équivalence et approximation

Complexité

Avec un vocabulaire de 65 000 mots :

- $65\,000^2 = 4\,225\,000\,000$ phrases de 2 mots possibles,
- $65\,000^3 = 2,74 \times 10^{14}$ phrases de 3 mots,

Classe d'équivalence

Regroupement des historiques en classe d'équivalence Φ

$$P(\boldsymbol{W}) \approx \prod_{i=1}^{T} P(w_i | \Phi(w_1, w_2, \ldots w_{i-1})) \quad (4)$$

"Tout l'art de la modélisation du langage consiste à déterminer Φ et une méthode pour estimer les probabilités associées"

F. Jelinek

Classe d'équivalence et approximation

Complexité

Avec un vocabulaire de 65 000 mots :

- $65\,000^2 = 4\,225\,000\,000$ phrases de 2 mots possibles,
- $65\,000^3 = 2,74 \times 10^{14}$ phrases de 3 mots,

Classe d'équivalence

Regroupement des historiques en classe d'équivalence Φ

$$P(\boldsymbol{W}) \approx \prod_{i=1}^{T} P(w_i | \Phi(w_1, w_2, \ldots w_{i-1})) \tag{4}$$

"Tout l'art de la modélisation du langage consiste à déterminer Φ et une méthode pour estimer les probabilités associées"

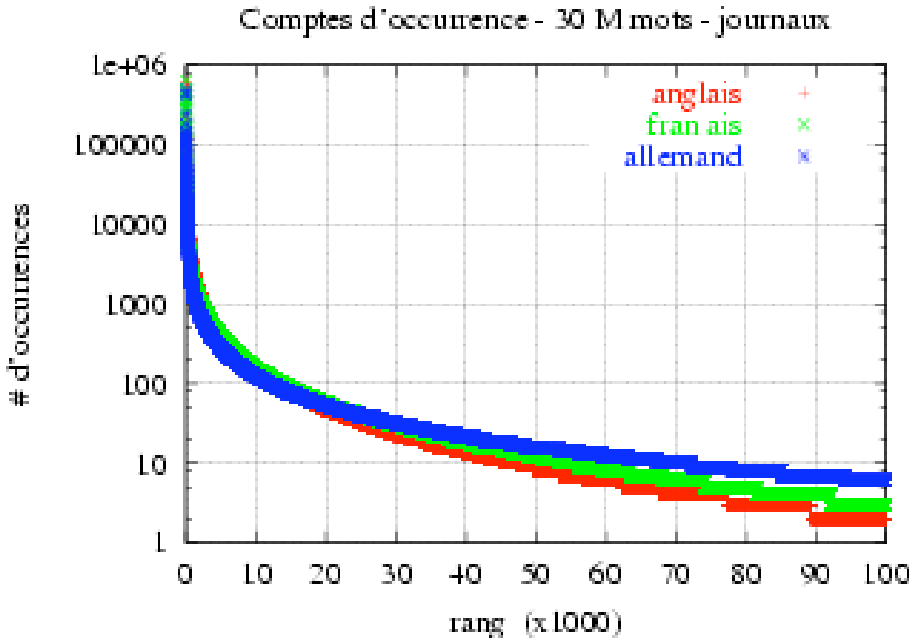
F. Jelinek

Évènements non observés

Loi de Zipf

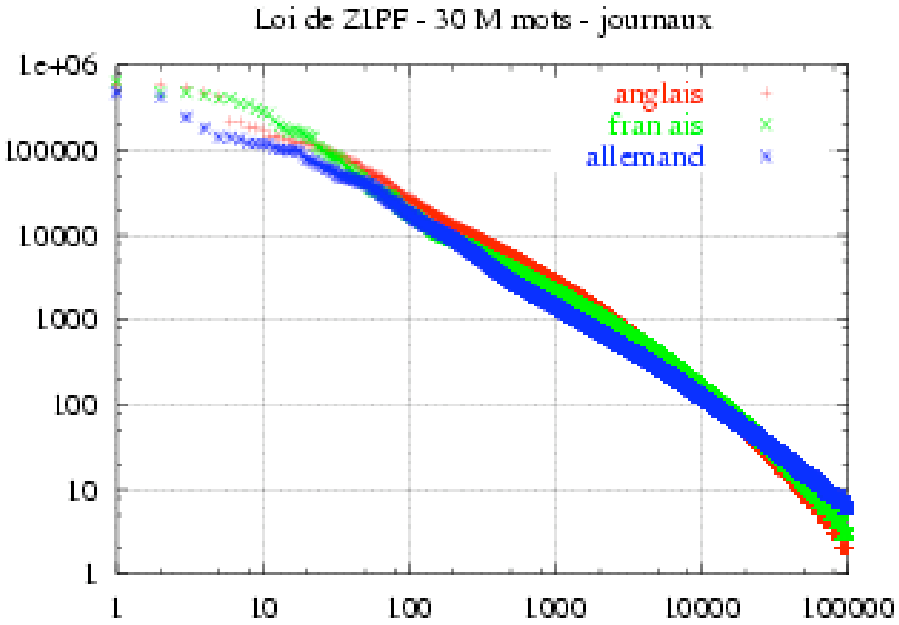
$$\text{Loi de Zipf : } f \approx \frac{K}{r},$$

Comptes d'occurrence triés par rang de fréquence obtenus sur des corpus de journaux de 30M de mots chacun.



fréquence *vs* rang

Pour les 100k mots les plus fréquents de chaque langue



fréquence de fréquence

Plan

Modèle *n*-gramme de mots

Modélisation du langage par une source markovienne d'ordre $n-1$

La probabilité d'émission d'un mot dépend exclusivement des $n-1$ précédents.

Décomposition d'une séquence de mots

$n=2$	bigramme	$P(\boldsymbol{W}) = P(w_1)\prod_{i=2}^{T} P(w_i \mid w_{i-1})$
$n=3$	trigramme	$P(\boldsymbol{W}) = P(w_1)P(w_2 \mid w_1)\prod_{t=3}^{T} P(w_i \mid w_{i-1}\, w_{i-2})$
n	n-gramme	$P(w_i \mid w_{i-1} \ldots w_{i-n+1})$

Conséquences

- $n-1$ mots suffiraient à prédire un mot.
- En pratique $n \leq 4$
- Classe d'équivalence Φ : $(w_1, w_2, \ldots w_{i-1}) \rightarrow (w_{i-n+1} \ldots w_{i-1})$

Caractéristiques des modèles *n*-gramme de mots

Une modélisation fondée sur les régularités du langage

- Structure du langage capturée implicitement sous forme d'une probabilité de succession de mots.
- Probabilité indépendante de la position dans la phrase (des mots spéciaux indiquent le début et la fin de phrase, `<s>`, `</s>`).
- Probabilités estimées à l'aide de grand corpus de textes

Hypothèse d'indépendance

- *quid* des phrases de plus de *n* mots
- *quid* des dépendance inter-phrases (ex : anaphore)

Estimation des probabilité

Estimateur du **Maximum de vraisemblance**

Unigramme

Unigramme : estimation de la probabilité d'un mot w_i :

$$P(w_i) = \frac{C(w_i)}{\sum_k C(w_k)} = \frac{C(w_i)}{\text{taille du corpus}}$$

n-gramme

estimation de la probabilité conditionnelle d'un mot w_i étant donné son historique h^{n-1} de $n-1$ mots précédents :

$$P(w_i|h^{n-1}) = \frac{C(h^{n-1}w_i)}{C(h^{n-1})}$$

dans le cas d'un bigramme $h^{n-1} = w_j$, le prédécesseur de w_i

Plan

Évaluation d'un ML : perplexité

Théorie de l'information

- La perplexité d'un texte mesurée avec un modèle de langage quantifie la diminution de l'entropie d'un texte due à l'utilisation de ce modèle.
- la capacité du ML à prédire les mots d'un texte "inconnu".
- Soit $\boldsymbol{W} = w_1 w_2 ... w_T$ une séquence de mots test $PP \stackrel{def}{=} P(\boldsymbol{W})^{-\frac{1}{T}}$

Un facteur de branchement

Interprétation comme facteur de branchement : l'utilisation du modèle revient à choisir entre PP mots équiprobables après chaque mot.

$$PP \leq N, \text{ la taille du vocabulaire}$$

$\Rightarrow$ Perplexité faible $\rightarrow$ bon ML !

Imperfection

Ne tient pas compte de la proximité phonétique des mots (gênant pour la RAP), dépendante du texte, ...

Perplexité...

Relation avec l'entropie croisée du modèle

$$
\begin{aligned}
PP &= 2^H \\
\text{avec } H &= \log \prod_{i=1}^{T} P(w_i)^{-\frac{1}{T}} \\
&= \sum_{i=1}^{T} \log P(w_i)^{-\frac{1}{T}} \\
&= \frac{1}{T} \sum_{i=1}^{T} -\log P(w_i)
\end{aligned}
$$

Procédure de calcul

- calculer la somme des logarithmes négatifs des probabilités N-grammes
- normaliser cette somme par T

Exemple de calcul

P(Le président François Holland a présenté ses voeux) = ??

2-grammes		3-grammes	
$P(\text{le} \mid < s >)$	1.3941	$P(\text{le} \mid < s >)$	1.3009
$P(\text{président} \mid \text{le})$	1.7206	$P(\text{président} \mid < s >, \text{le})$	1.3844
$P(\text{François} \mid \text{président})$	2.4011	$P(\text{François} \mid \text{le}, \text{président})$	2.2343
$P(\text{Holland} \mid \text{François})$	0.3444	$P(\text{Holland} \mid \text{président}, \text{François})$	0.1158
$P(\text{a} \mid \text{Holland})$	1.0458	$P(\text{a} \mid \text{François}, \text{Holland})$	0.9839
$P(\text{présenté} \mid \text{a})$	2.7520	$P(\text{présenté} \mid \text{Holland}, \text{a})$	2.5205
$P(\text{ses} \mid \text{présenté})$	2.0150	$P(\text{ses} \mid \text{a}, \text{présenté})$	1.5563
$P(\text{voeux} \mid \text{ses})$	2.5941	$P(\text{voeux} \mid \text{présenté}, \text{ses})$	1.7149
$P(< /s > \mid \text{voeux})$	1.4140	$P(< /s > \mid \text{ses}, \text{voeux})$	1.2823
=	15.6819	=	13.0930
$\Rightarrow$ PP =	55.2625	$\Rightarrow$ PP =	28.4956

Un modèle de language = un ensemble de classifieur

Pour un historique donné ou $\Phi(w_1, w_2, ...w_{i-1})$

- Inférer le mot suivant w_i connaissant $\Phi(w_1, w_2, ...w_{i-1})$.
- Une distribution sur $\boldsymbol{V}$ pour un $\Phi(w_1, w_2, ...w_{i-1})$ donné.
- En pratique: une multinomiale par historique

Un modèle *n*-gramme =

- Un regroupement de multinomiales.
- Il faut lier les paramètres.

Plan

Plan

Problèmes : mots hors vocabulaire

Mots hors vocabulaire (*Out of Vocabulary Words, OOV*)

- lorsqu'un mot est hors vocabulaire, sa probabilité serait nulle et donc la perplexité $\rightarrow \infty$
- Lors de l'apprentissage il faut minimiser le taux de OOV et associer une probabilité à tous les mots OOV
- $\rightarrow < UNK >$

Différences entre les langues

- l'anglais et le français sont comparables.
- l'allemand utilise beaucoup de mots composés $\rightarrow$ taux d'OOV plus élevé.

Problèmes : observations des *n*-grammes

- Quantité de données d'apprentissage : il n'y a jamais assez de données textuelles pour estimer toutes ces probabilités
- ex. 1 000 mots $\rightarrow 10^6$ bigrammes, 10^9 trigrammes
- Hypothèse markovienne insuffisante en pratique
- Exemple : trigramme

$$P(w_i|w_j, w_l) = \frac{C(w_j w_l w_i)}{C(w_j w_l)}$$

- Problèmes $P(w_i|w_j, w_l) = 0$ si $C(w_j w_l w_i) = 0$ et $C(w_j w_l) \neq 0$

  $P(w_i|w_j, w_l) = \infty$ si $C(w_j w_l) = 0$

Évènements non observés

Différentes raisons

- séquences non admises par la syntaxe de la langue
  ex. *ils part tôt* (typiquement des accords non-ambigus en genre et en nombre)
- mauvaise raison : séquences absentes du corpus, mais faisant partie de la langue

Solutions :

- augmenter la taille du corpus d'apprentissage;
- un modèle capable de généraliser ses connaissances;
- attribuer une probabilité (faible) aux événements non observés :

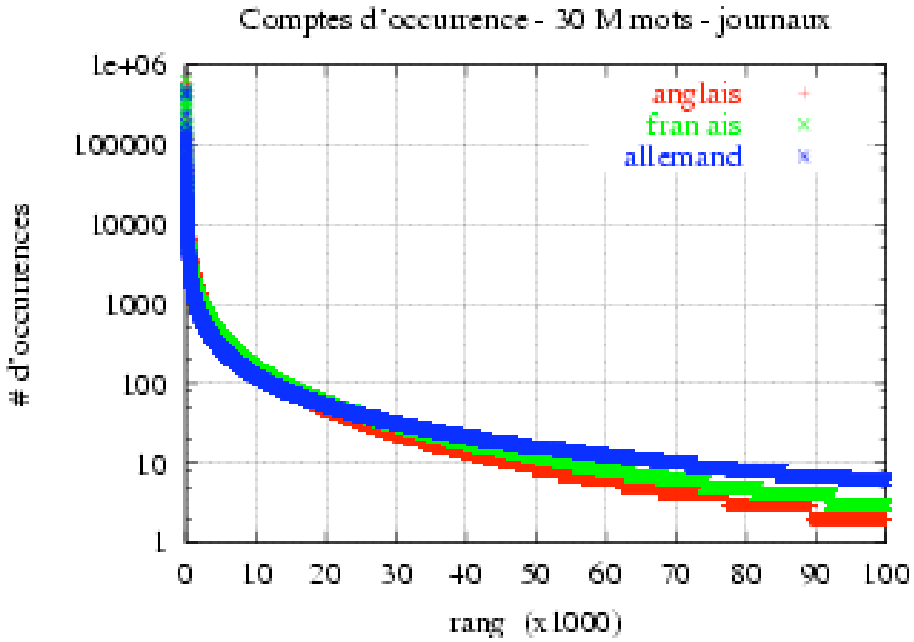
$$P(w_i|h) \geq \epsilon > 0 \quad \forall i, \ \forall h$$

Évènements non observés

Loi de Zipf

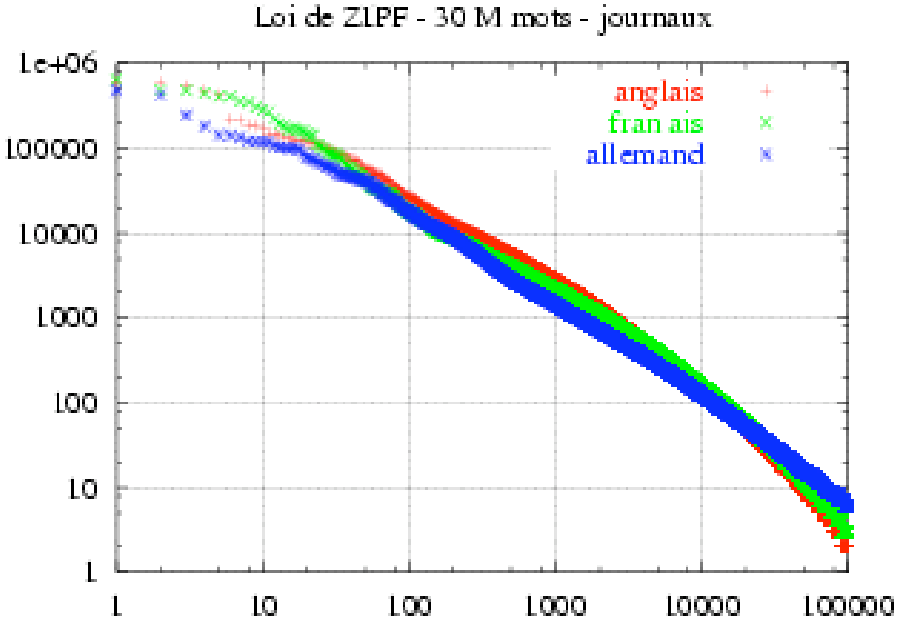
$$\text{Loi de Zipf : } f \approx \frac{K}{r},$$

Comptes d'occurrence triés par rang de fréquence obtenus sur des corpus de journaux de 30M de mots chacun.



fréquence *vs* rang

Pour les 100k mots les plus fréquents de chaque langue



fréquence de fréquence

Prélèvement

Prélever une **masse de probabilité** D aux événements observés,

$$P^-(w_i|h^n) = (1 - \delta(w_i, h^n))P_{MV}(w_i|h^n)$$

puis la redistribuer sur les événements n_0 non-observés.

Techniques de prélèvement :

- prélèvement constant : δ est une constante
- **prélèvement absolu (***absolute discounting ou Knesser-Ney***)** : $\delta(h) = \frac{1}{f(h)}$
- prélèvement Good-Turing et Katz : dépend de la fréquence des événements observés et de la fréquence des fréquences
- prélèvement Witten-Bell : on rajoute un poids au dénominateur, lié au nombre d'événements distincts observés.
- Se référer à [**?**] (en version rapport technique)

méthodes de lissage

Le lissage combine le prélèvement et la redistribution.

Lissage et généralisation

Le lissage opère un *saut inductif*: généralise le corpus fini à un langage infini;
$\Rightarrow$ *Toute phrase a une probabilité non nulle*.

Les classiques

- Combinaison linéaire (*interpolation*) de plusieurs modèles;
- Le repli (*back-off*)
- Introduction d'*a priori* sur les paramètres;
- Construction de classes de mots $\Rightarrow$ classes d'histoires;
- ...

Redistribution - Interpolation linéaire

Comment faire un compromis

- Les modèles les plus simples (unigramme) sont les mieux estimés.
- L'ordre du modèle augmente sa capacité de prédiction.

Combinaison linéaire de modèles de complexité croissante selon

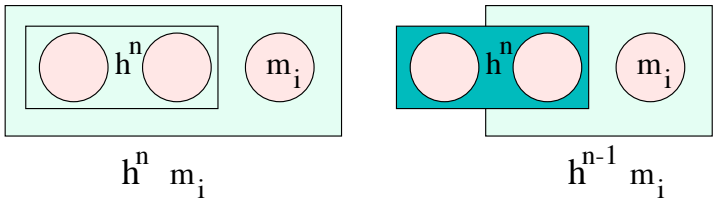
$$P_l(w_i \mid h^n) = \lambda(w_i, h^n) P^-(w_i \mid h^n) + (1 - \lambda(w_i, h^n)) P_l(w_i \mid h^{n-1})$$

Le coefficient λ est de manière générique une fonction du mot et de l'historique

Redistribution - Repli

Principe : exploiter les historiques d'ordre plus faible

non observé **approximation**



- Technique de repli (back-off) $P^-() \rightarrow \tilde{P}()$

$$\tilde{P}(w_i|h^n) = \begin{cases} P^-(w_i|h^n) & \text{si } C(h^n w_i) > 0 \\ \alpha(h^n) P^-(w_i|h^{n-1}) & \text{si } C(h^n w_i) = 0 \end{cases}$$

- $\alpha(h^n)$: coefficient de repli (*back-off*) déterminé pour remplir la condition de normalisation des probabilités conditionnelles

Le plus simple, le prélèvement constant ou additif

Définition

$$P_{add}(w_i|h^n) = \frac{c(w_i, h^n) + \alpha}{\alpha|V| + \sum_w c(w_i, h^n)}$$

La constante est telle que $0 < \alpha < 1$

- Cette méthode est en général peu efficace.
- Mieux vaut prélever en fonction du mot ou de l'historique

Aussi simple : Jelinek-Mercer

Définition récursive de l'estimation via l'interpolation linéaire.

Définition

$$P_{Jel}(w_i \mid h^n) = \lambda(w_i, h^n) P_{MV}(w_i \mid h^n) + (1 - \lambda(w_i, h^n)) P_{Jel}(w_i \mid h^{n-1})$$

Estimation des paramètres

doit se faire sur des données de validation ou :

- Held-out-data
- Deleted Interpolation

Katz / Good-Turing

Idée de départ

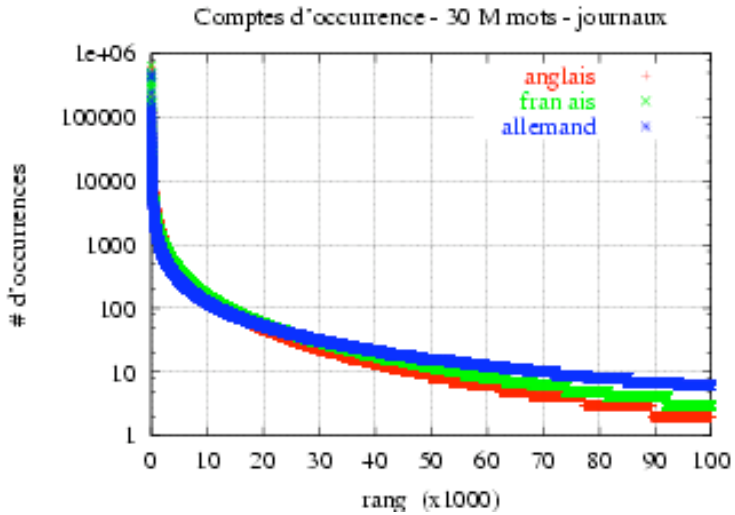
L'estimation MV surestime les évènements rares → Correction de la fréquence des événements.

Prélèvement

Soit n_r le nombre de n-grams apparaissant r fois :

$$r^* = disc(r) = (r+1)\frac{n_{r+1}}{n_r}$$

Prise en compte de la loi de Zipf



Les fréquences de Good-Turing

Redistribution

$$P_K(w_i|h^n) = \begin{cases} P^-(w_i|h^n) = \frac{disc(C(h^n w_i))}{C(h^n)} & \text{si } C(h^n w_i) > 0 \\ \alpha(h^n) P^-(w_i|h^{n-1}) & \text{si } C(h^n w_i) = 0 \end{cases}$$

r	n_r	r^*	r	n_r	r^*
1	1963237	0.22	11	3341	9.69
2	211420	1.01	12	2697	10.89
3	71258	1.95	13	2259	11.86
4	34795	2.94	14	1914	12.86
5	20471	3.87	15	1641	13.81
6	13215	4.83	16	1416	15.03
7	9109	5.89	17	1252	16.55
8	6709	5280	18	1151	16.84
9	5280	7.82	19	1020	16.20
10	4127	8.91	20	826	19.78

Chaque évènement non-observé a un compte GT $\approx 10^{-6}$.

Inconvénient de la méthode de Katz

Surestimation des probabilités

- $C(en) = 10000$, $(le) = 10000$, $C(appel) = 100$,
  $C(en\ le) = C(en\ appel) = 0$

$\rightarrow$ $P(en\ le) >> P(en\ appel)$!!

- Probabilité du bigramme inobservé *on Frisco*
  $P_K(Frisco/on) = \alpha(on) P_K(Frisco)$

  Comme *Frisco* est très fréquent, $P_K(Frisco/on)$ est élevée
  Pourtant *Frisco*, même si très fréquent, apparaît dans peu de contexte.

Idée : prendre en compte la propension du mot à se combiner à gauche

Le mot apparait-il dans de nombreux contextes ou est-il spécifique ?

Witten-Bell

Interpolation avec l'ordre inférieur

$$P_{wb}(w_i|h^n) = \lambda_{h^n} P_{MV}(w_i|h^n) + (1 - \lambda_{h^n}) P_{wb}(w_i|h^{n-1})$$

Le coefficient d'interpolation est fonction de l'historique et de "sa spécificité", soit $|\{w\}, C(h^n w) > 0|$:

$$\lambda_{h^n} = \frac{|\{w\}, C(h^n w) > 0|}{|\{w\}, C(h^n w) > 0| + \sum_w c(h^n, w)}$$

Knesser-Ney

$$\tilde{P}(w_i|h^n) = \begin{cases} P^-(w_i|h^n) = \frac{C(h^n w_i) - D}{C(h^n)} & \text{si } C(h^n w_i) > 0 \\ \alpha(h^n) \frac{|\{v^n\}, C(v^n w_i) > 0|}{\sum_w |\{v^n\}, C(v^n w) > 0} & \text{si } C(h^n w_i) = 0 \end{cases}$$

- D est une constante (prélèvement absolu)
- $|\{v^n\}, C(v^n w_i) > 0|$ est le nombre d'historique possible pour w_i
- Raffinement : D est optimisé indépendemment pour les n-gram apparaissant 1,2 et 3 fois

Plan

Normalisation des corpus d'entraînement

La normalisation des textes est un compromis entre :

Couverture lexicale

Obtenir une meilleure couverture lexicale et un meilleur apprentissage → réduction du nombre de mots

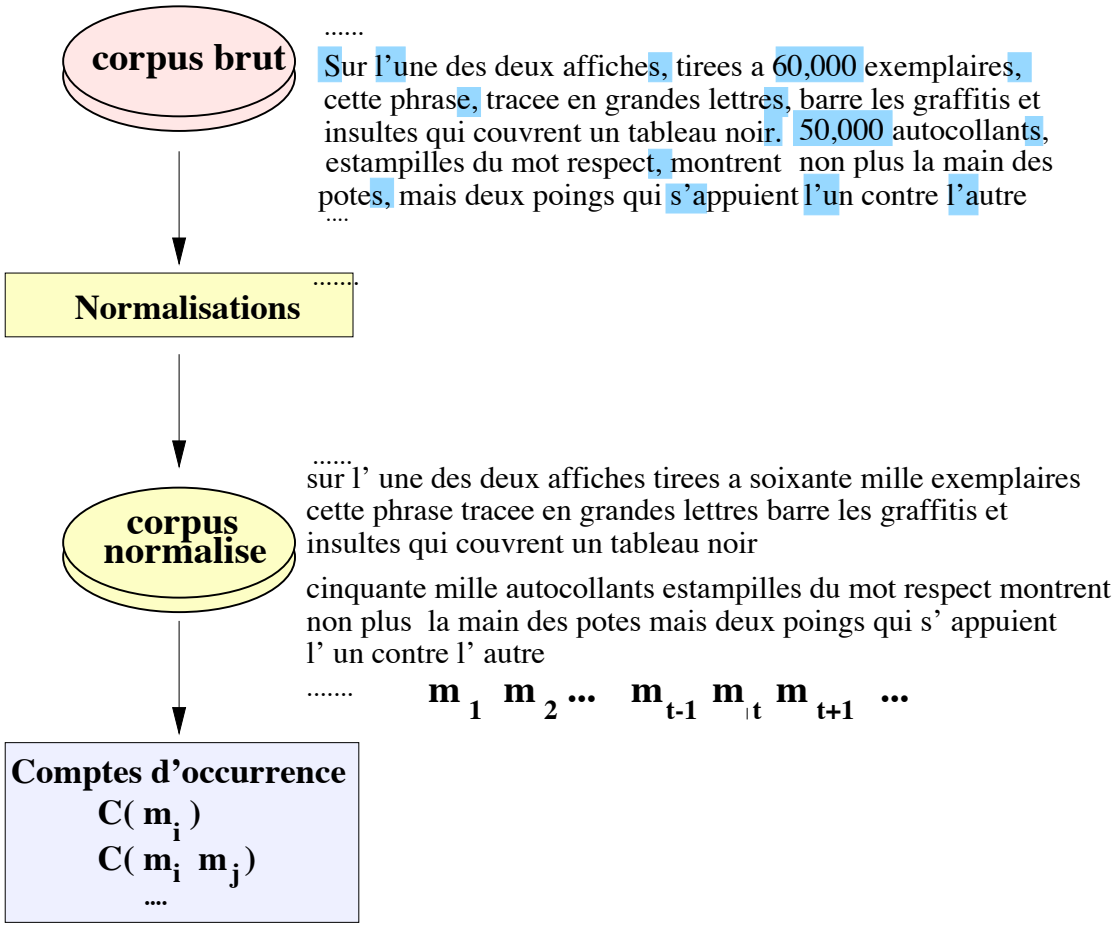
- conversion des chiffres en mots
- éclatement des sigles
- uniformiser l'écriture des unités

Capacité discriminante

Discrimination du modèle de langue → conservation de toutes les distinctions

- conserver la capitalisation (français) et les acronymes, ex : *Roman* ou *roman*
- écritures alternatives et contraction, ex. : *we'll* ou *we will*

Chaîne de traitement



- A partir d'un corpus (suites de mots $w_1 \ldots w_{t-1} w_t w_{t+1} \ldots w_T$)
- on détermine les mots distincts (vocabulaire w_i, $i = 1, \ldots I$) et
- les comptes d'occurrences de suites de mots de longueur 1, 2, 3 ...

Choix des corpus d'entraînement

- Corpus en relation avec la tâche $\rightarrow$ caractérisation de la tâche ?
  ex. : transcription d'émissions télévisées d'actualité
  - Utilisation de transcription d'émissions télévisées d'actualité
    mais quantité insuffisante
  - Utilisation de textes de la presse écrite
    mais langue écrite $\neq$ parole spontanée

  $\rightarrow$ par ex. ajouter des hésitations, des respirations
- Quelle actualité, quelle langue ? $\rightarrow$ époque des textes utilisés
  - anciens $\rightarrow$ mots d'intérêt général
  - récents $\rightarrow$ surtout pour les noms propres

Modèle de langue en français

Répartition des données d'entraînement

Source	*Moyenne annuelle en million de mots*	*Période couverte*	*Total en million de mots*
Transcriptions	0,3	1994-1999	1,6
Service de presse	24,2	1997,1998,2000	72,7
Agence de presse	22,3	1994-1996	66,8
Le Monde	21,4	1987-1998	257,4
Le Monde Diplo.	1,0	1990-1996	6,7

Modèle de langue en français

Répartition des données d'entraînement

Source	*Moyenne annuelle en million de mots*	*Période couverte*	*Total en million de mots*
Transcriptions	0,3	1994-1999	1,6
Service de presse	24,2	1997,1998,2000	72,7
Agence de presse	22,3	1994-1996	66,8
Le Monde	21,4	1987-1998	257,4
Le Monde Diplo.	1,0	1990-1996	6,7

Modèle de langue en français

Répartition des données d'entraînement

Source	*Moyenne annuelle en million de mots*	*Période couverte*	*Total en million de mots*
Transcriptions	0,3	1994-1999	1,6
Service de presse	24,2	1997,1998,2000	72,7
Agence de presse	22,3	1994-1996	66,8
Le Monde	21,4	1987-1998	257,4
Le Monde Diplo.	1,0	1990-1996	6,7

Modèle de langue en français

Répartition des données d'entraînement

Source	*Moyenne annuelle en million de mots*	*Période couverte*	*Total en million de mots*
Transcriptions	0,3	1994-1999	1,6
Service de presse	24,2	1997,1998,2000	72,7
Agence de presse	22,3	1994-1996	66,8
Le Monde	21,4	1987-1998	257,4
Le Monde Diplo.	1,0	1990-1996	6,7

Modèle de langue en français

Répartition des données d'entraînement

Source	*Moyenne annuelle en million de mots*	*Période couverte*	*Total en million de mots*
Transcriptions	0,3	1994-1999	1,6
Service de presse	24,2	1997,1998,2000	72,7
Agence de presse	22,3	1994-1996	66,8
Le Monde	21,4	1987-1998	257,4
Le Monde Diplo.	1,0	1990-1996	6,7

Modèle de langue en français

Répartition des données d'entraînement

Source	*Moyenne annuelle en million de mots*	*Période couverte*	*Total en million de mots*
Transcriptions	0,3	1994-1999	1,6
Service de presse	24,2	1997,1998,2000	72,7
Agence de presse	22,3	1994-1996	66,8
Le Monde	21,4	1987-1998	257,4
Le Monde Diplo.	1,0	1990-1996	6,7

Éviter la dilution : Un modèle par source est estimé (transcription, service de presse, presse écrite)

Construction du modèle de langue de référence

- Interpolation linéaire des trois modèles :

$$P_{interpol}(w|h) = \sum_{i=1}^{3} \lambda_i P_i(w|h), \text{ avec } \sum_{i=1}^{3} \lambda_i = 1.$$

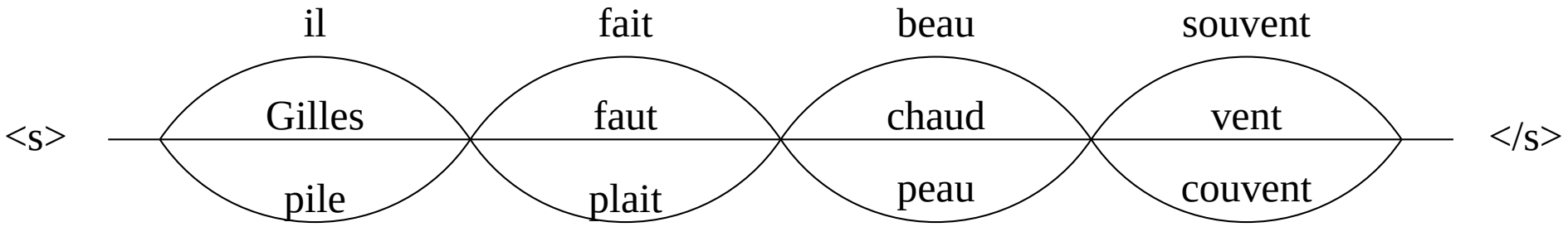
- les (λ_i) calculés de manière à minimiser la perplexité d'un texte de développement T :

$$ppx(T) = 2^{\mathcal{L}(T)} \text{ avec } \mathcal{L}(T) = \frac{1}{n} \sum_{j=1}^{n} \log_2 P(w_j|h_j)$$

$\Rightarrow$ 15 millions de bigrammes, 13 millions de trigrammes, 10 millions de quadrigrammes

Trouver la phrase la plus fréquente

Considérons le graphe suivant :



Comment déterminer la phrase (suite de mots) la plus probable en utilisant un modèle de langue bi- ou trigrammes ?

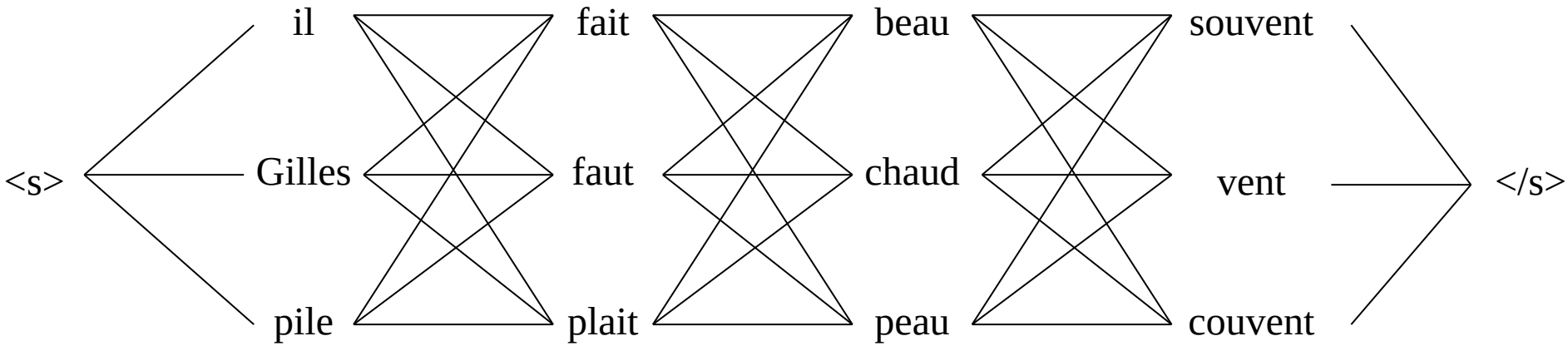
Tester toutes les phrases : $3^4 \times 5$ bigrammes ≈ 400 évaluations

$\Rightarrow$ programmation dynamique

$\Rightarrow$ algorithme ”classique” de parcours de graphe

Algorithmes...

Exemple avec des bigrammes : créer un graphe dont les noeuds sont les mots et dont les arcs sont les probabilités des bigrammes correspondants



$\rightarrow$ chercher le chemin maximal par programmation dynamique
(33 évaluations)
Comment faire en cas de tri-grammes ?
$\Rightarrow$ graphe de couples de mots

Conclusion sur ...

la modélisation linguistique en Reconnaissance automatique de la parole continue grand vocabulaire.

l'état de l'art :

- modèle résultant de l'interpolation de ML *n*-gramme de mots, données d'entraînement $>$ 1 milliard de mots. Smoothing : *Knesser-Ney modified*
- un ML *n*-gramme de classe peut-être interpolé avec le ML *n*-gramme de mots

Perspectives

- Put the language back into the language modeling.
- Les performances atteintes commencent à être suffisante pour envisager des applications nouvelles comme l'indexation automatique de documents audiovisuels (cours 6).
- Nouvelles perspectives de recherche sur l'extraction de connaissance et fouille de données audio(visuelles).

Bibliographie ”courte”

Les références

- Frederick Jelinek, ”Statistical Methods for Speech Recognition”, The MIT Press, 2000
- Christopher D. Manning and Hinrich Schütze, ”Foundations of Statistical Natural Language Processing”, The MIT Press, 1999
- Ronald Rosenfeld, ”Two decades of statistical language modeling: Where do we go from here ? ”, Proceedings of the IEEE, 88(8), 2000.
- Jean-Luc Gauvain and L. Lamel and G. Adda, “The LIMSI Broadcast News Transcription System”, Speech Communication, 37, 2002.

Do it yourself !

- www.speech.sri.com/projects/srilm/
- http://kheafield.com/code/kenlm/