Programmation de Modèles Linguistiques (I), partie 2 (L5SOPROG L3 SDL)

Crédits : Karën Fort (pépites pour le TAL)

Caroline Koudoro-Parfait et Gaël Lejeune
caroline.parfait@sorbonne-universite.fr
gael.lejeune@sorbonne-universite.fr

2024-2025

Observatoire des Textes des Idées et des Corpus - Obtic,
Sorbonne Center for Artificial Intelligence - SCAI,
Sens Textes Informatiques Histoire - STIH EA 4509, Sorbonne Université

Plan de la présentation

1. Les solutions

Les solutions

Les solutions

- Typologie des techniques
- Identification de la langue

Des techniques variées

- systèmes à base de **règles**
  - définies par l'humain (linguistes)
  - entrées manuellement
- systèmes basés sur les **données**
  - apprentissage supervisé ou non supervisé
  - à partir d'exemples (rédigés et/ou annotés par des humains)
  - algorithmes (pensés par des humains)

$\rightarrow$ et tous les intermédiaires/mélanges possibles (approches hybrides)

Des techniques variées

- systèmes à base de **règles**
  - définies par l'humain (linguistes)
  - entrées manuellement
- systèmes basés sur les **données**
  - apprentissage supervisé ou non supervisé
  - à partir d'exemples (rédigés et/ou annotés par des humains)
  - algorithmes (pensés par des humains)

$\rightarrow$ et tous les intermédiaires/mélanges possibles (approches hybrides)

- Traduction automatique :
  - analyse linguistique + traitement statistiques
  - génération statistique + redressement linguistique

Les solutions

Typologie des techniques

Identification de la langue

Exercice : identifier la langue d'un texte

Trouver au moins **2** algorithmes permettant d'identifier la langue d'un texte

Conseil : connaître les langues et leurs caractéristiques

Rang (r)	Mot	Fréquence (f)	$\frac{f}{r}$
1	*the*	69 971	69 971
2	*of*	36 412	34 986
3	*and*	28 853	23 324
...	...	...	...
20	*I*	5 164	3 499

Loi de Zipf sur le Brown corpus

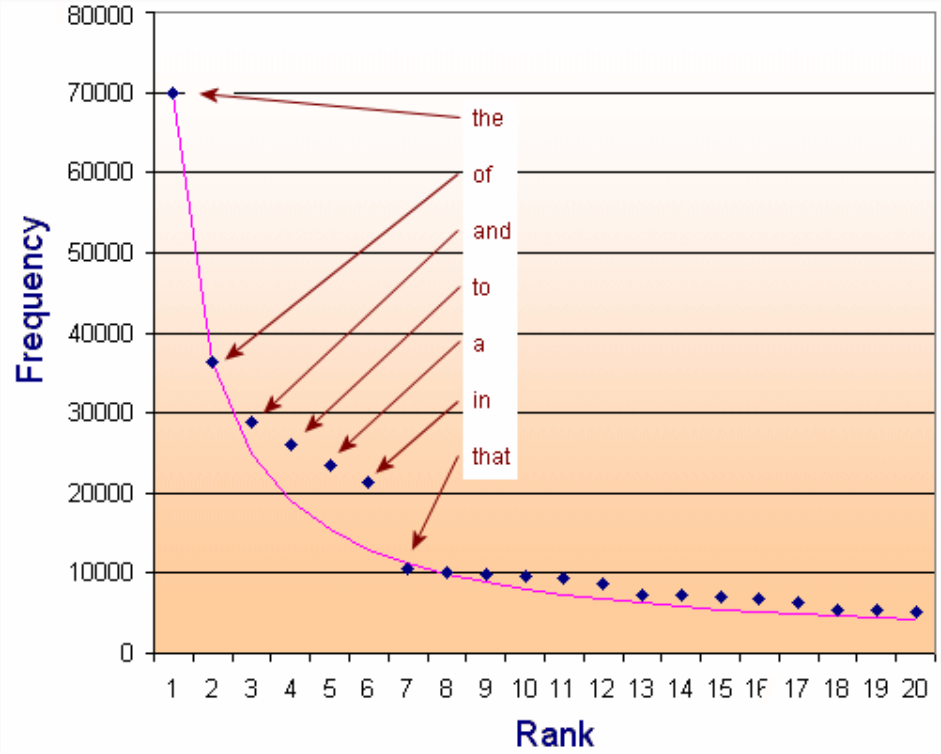


Figure 1: Données très proches de l'attendu, surtout sur la longue traîne

Loi de Zipf sur le Brown corpus

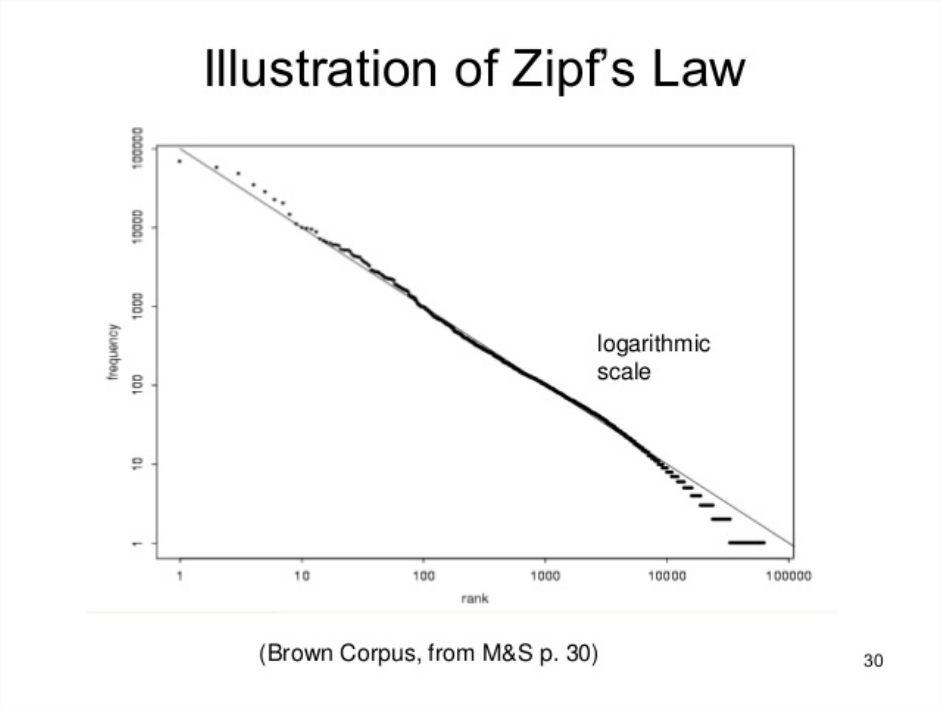


Figure 2: Validité plus marquante encore en échelle logarithmique

Identifier la langue : solution 1

Méthode des *short words* / *frequent words* :

- Liste de "mots outils" (mots grammaticaux, "petits" mots) pour chaque langue
- Compter les occurrences de ces mots outils dans le texte
- Comparer avec des listes de référence

Implantation rapide (POC)

Données : corpus parallèle de l'Union Européenne (22 langues)

- Découpage en deux parties (entraînement et test)
- Entraînement : extraction d'un modèle de langue (les n mots les plus fréquents) à partir de tous les textes de chaque langue

Implantation rapide (POC)

Données : corpus parallèle de l'Union Européenne (22 langues)

- Découpage en deux parties (entraînement et test)
- Entraînement : extraction d'un modèle de langue (les n mots les plus fréquents) à partir de tous les textes de chaque langue
- Test, pour chaque texte :
  - calcul de l'intersection en mots
  - on prend la plus grande $\rightarrow$ prédiction

Les modèles

lg	#1	#2	#3	#4	#5
bg	на (12593)	за (5657)	и (5529)	в (3919)	от (3474)
cs	a (5510)	v (3378)	na (2424)	se (1955)	pro (1668)
da	og (5435)	i (4542)	at (4147)	af (3682)	for (3636)
de	der (5867)	die (5604)	und (5155)	in (2747)	für (2256)
en	**the (9547)**	and (5692)	of (5430)	to (4787)	in (3667)
es	de (16556)	la (8571)	en (5096)	y (5048)	los (4721)
et	ja (4295)	on (2746)	Euroopa (1658)	et (1240)	ning (102)
fi	ja (4952)	on (2623)	Euroopan (985)	EU :n (898)	että (875)
fr	**de (11801)**	la (6466)	et (5177)	les (4999)	des (4821)
hu	a (9824)	az (4956)	és (4327)	A (2509)	hogy (17(
it	di (7617)	e (4838)	in (2987)	la (2958)	per (2746)
lt	ir (4984)	Europos (1645)	kad (1311)	– (1293)	ES (1247)
lv	un (5028)	ir (2448)	par (1658)	Eiropas (1473)	ES (1261)
mt	u (5234)	li (4557)	ta’ (2960)	ta’ (1554)	biex (123)
nl	de (11253)	van (7093)	en (5167)	het (3986)	in (3687)
pl	w (5750)	i (3799)	na (2844)	z (1986)	do (1890)
pt	de (10488)	a (6684)	e (5153)	da (3785)	o (2983)
ro	de (10094)	în (5478)	și (5020)	a (4710)	la (2816)

L'application

Référence	Préd 1	Préd 2	Préd 3
cs	sk (2)	cs (2)	sl (1)
cs	sk (4)	cs (3)	pt (2)
cs	sk (4)	cs (4)	sl (2)
cs	sk (5)	cs (5)	sl (3)
cs	sk (5)	cs (5)	sl (3)
cs	sk (6)	cs (6)	sl (3)
cs	sk (6)	cs (6)	sl (4)
cs	sl (3)	sk (3)	cs (3)
et	fi (2)	et (2)	en (1)
et	fi (2)	et (2)	en (1)
et	fi (2)	et (2)	en (1)
et	fi (2)	et (2)	en (1)
et	fi (2)	et (2)	en (2)
et	fi (3)	et (3)	en (3)
bg	en (8)	fi (3)	et (3)
cs	en (8)	fi (3)	et (3)
da	en (8)	fi (3)	et (3)
de	en (8)	fi (3)	et (3)

Identifier la langue : solution 2

Méthode des *trigrammes* :

- Rechercher la probabilité qu'un caractère C_i apparaisse après les deux précédents dans la langue l :

$$P(C_i | C_{i-2} : C_{i-1}, l)$$

- Calculer la probabilité résultante pour chaque langue, pour l'ensemble du texte :

$$\prod_{i=1}^{i=n} P(C_i | C_{i-2} : C_{i-1}, l)$$

Dans quelle langue est ce texte :

Roiajr oj earoij reoa o eo ao aeoi oj aeroij aoeir eoaj

Sachant que :

- L1 : P(i,ro, L1)=0,3 ; P(i,eo, L1)=0,2 ; P(i,oe, L1)=0,3
- L2 : P(i,ro, L2)=0,8 ;P(i,eo, L2)=0,2 ; P(i,oe, L2)=0,3

Dans quelle langue est ce texte :

Roiajr oj earoij reoa o eo ao aeoi oj aeroij aoeir eoaj

Sachant que :

- L1 : P(i,ro, L1)=0,3 ; P(i,eo, L1)=0,2 ; P(i,oe, L1)=0,3
- L2 : P(i,ro, L2)=0,8 ;P(i,eo, L2)=0,2 ; P(i,oe, L2)=0,3

- P(c,L1)=0,3*0,3*0,2*0,3*0,3 = 0,00162
- P(c,L2)=0,8*0,8*0,2*0,8*0,3= 0,03072

Autres Modèles : 3-grammes de caractère

lg	#1	#2	#3	#4	#5	6
bg	_на (12863)	на_ (11886)	ите (9741)	_за (6523)	та_ (6271)	_п
da	er_ (14032)	en_ (9306)	for (8681)	_de (8165)	_fo (7199)	et_
en	_th (13006)	the (11879)	he_ (11177)	ion (8614)	and (6666)	_in
es	_de (20787)	de_ (16648)	os_ (13741)	_la (11721)	as_ (9391)	es_
et	mis (6513)	se_ (5245)	ise (4791)	ja_ (4568)	_ja (4563)	ust
fi	en_ (11551)	ist (6937)	an_ (6291)	sta (6028)	ja_ (5459)	ta_
fr	es_ (21305)	_de (17707)	de_ (12042)	ion (11016)	ent (9673)	_le
hu	_a_ (8998)	_az (5594)	és_ (4906)	az_ (4712)	_sz (4534)	_és
it	ion (9886)	_di (9647)	_de (9207)	di_ (7761)	re_ (7434)	to_
lt	os_ (9469)	_pa (6289)	_ir (4924)	ir_ (4770)	ti_ (4449)	_pr
lv	as_ (11209)	_pa (5859)	_un (5018)	un_ (4714)	s_p (4065)	iem
mt	_ta (14740)	tal (7746)	al- (7613)	li_ (7590)	jon (6872)	oni
nl	en_ (25906)	de_ (13221)	_de (12334)	an_ (9452)	van (7780)	n_d
pl	nie (7586)	ch_ (7460)	_pr (7326)	ie_ (7261)	ych (5844)	_po
pt	_de (13126)	os_ (12968)	de_ (11863)	as_ (9777)	ent (7858)	ão_
sk	_pr (8264)	ch_ (5970)	_po (5275)	_na (4609)	ie_ (4094)	ých
sl	_pr (7414)	_po (7173)	je_ (7010)	_in (6385)	_za (6004)	_na
ro	_de (12515)	de_ (10232)	are (8296)	_în (7364)	re_ (7350)	le_

Bilan : ça marche !

Plus de 96% de bonne prédiction sur 22 langues, `langid.py` fait encore mieux. Plus rapide et plus efficace que l'humain.

Mais pourquoi ?

Bilan : ça marche !

Plus de 96% de bonne prédiction sur 22 langues, `langid.py` fait encore mieux. Plus rapide et plus efficace que l'humain.

Mais pourquoi ?

- Des données disponibles
- Une tâche facile à définir (classification)
- Et facile à évaluer

Bilan : ça marche !

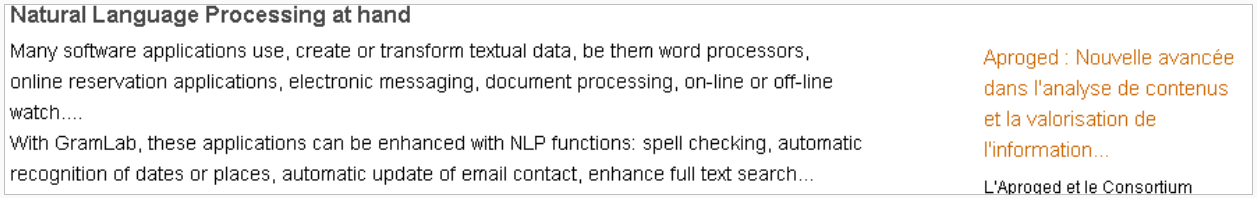
Plus de 96% de bonne prédiction sur 22 langues, `langid.py` fait encore mieux. Plus rapide et plus efficace que l'humain.

Mais pourquoi ?

- Des données disponibles
- Une tâche facile à définir (classification)
- Et facile à évaluer
- Une théorie linguistique bien stable . . .
- . . . et facile à rendre calculable

Limitations

- Longueur du texte (5 mots mini.)
- textes multilingues :



- mais aussi des pièges et du contexte :
  - Barack Obama $\rightarrow$ italien
  - Nicolas Sarkozy $\rightarrow$ polonais
  - Barack Obama and Nicolas Sarkozy $\rightarrow$ anglais
  - camping caravaning, trekking

Identifier la langue : solutions 3 et 4

- Identifier l'encodage
- Regarder les méta-données

Limitations

- Identification non-triviale
- Remplissage des méta-données (ex [Html)

Premier bilan des obstacles

C'est à vous

Premier bilan des obstacles

C'est à vous

- Quantité de données ;
- Qualité des données ;
- Qualité des méta-données ;
- Définition des tâches.