Language Models

J. Savoy
Université de Neuchâtel

Une approche probabiliste pour des applications en langue naturelle

1

Language Models

- A statistical view of the language
  Opposed to a logical approach
- Estimate the probability of occurrence of single forms, or *n*-grams of such forms (words, letters)
- Can we find general laws governing the word distribution?
- Are words used randomly?
- Does the word distribution differ from one author to the other? (stylometry)
- Language models for speech recognition, information retrieval, spelling correction, language identification, …

2

What is a Word?

- Select the word as unit of measurement
- What is a word?
  Richard Brown is painting in New York.
  I'll send you Luca's book.
  C|net, Micro$oft, ;-)
  l'école, d'aujourd'hui, le chemin de fer
- Other possibilities
  lemma (entry in the dictionary, dogs -> dog)
- Example: *I saw a man with a saw*
  Count
  7 word *tokens (forme)*
  5 word *types (vocable)* Vocabulary = {I, saw, a, man, with}

3

Word Frequency

The most frequent word types f(ω)

With
|V| = 7,792
for J. McCain
|V| = 7,663
for B. Obama
(2008)

the number of distinct types (or vocabulary size)

	McCain'08		Obama'08	
Rank	**Word**	**f(ω)**	**Word**	**f(ω)**
1	the	7759	the	13027
2	and	6157	and	10950
3	to	5413	to	9072
4	of	4773	that	7446
5	in	3137	of	6985
6	a	2940	we	6203
7	I	2345	a	5562
8	that	2243	in	5340
9	we	2160	is	4986
10	for	1762	I	4216

4

Word Frequency
Brown Corpus

Collected in 1961

A real sample

1,014,312 tokens

Given by lemmas (e.g., "be" = "is", "was", "be", "were", etc.)

Rank	Word	Freq.	%
1	the	69975	6.90%
2	be	39175	3.86%
3	of	36432	3.59%
4	and	28872	2.85%
5	to	26190	2.58%
6	a	23073	2.28%
7	in	20870	2.06%
8	he	19427	1.92%
9	have	12458	1.23%
10	it	10942	1.08%

5

Rank	Brown		US		
1	the	6.90%	the	4.69%	4.69%
2	be	3.86%	be	3.81%	8.50%
3	of	3.59%	and	3.78%	12.28%
4	and	2.85%	to	3.30%	15.58%
5	to	2.58%	of	2.61%	18.19%
6	a	2.28%	that	2.17%	20.36%
7	in	2.06%	a	1.95%	22.31%
8	**he**	1.92%	in	1.88%	24.19%
9	have	1.23%	**we**	1.85%	26.04%
10	it	1.08%	**I**	1.50%	27.54%
11	**that**	1.05%	have	1.36%	28.90%
12	for	0.89%	not	1.19%	30.09%
13	not	0.87%	for	1.18%	31.27%
14	I	0.83%	our	1.10%	32.37%
15	they	0.82%	it	1.01%	33.38%
16	with	0.72%	will	0.98%	34.36%
17	on	0.61%	this	0.85%	35.21%
18	**she**	0.60%	you	0.68%	35.89%

With 12 word-types, we cover 30% of all texts

6

Zipf's Law

- More a regularity than a strict law
- The frequency (of a word type) ($f(\omega)$) is related to the inverse of its rank (z) (with α = 1 for Zipf)
- We could use the absolute frequency ($f(\omega)$) of the relative frequency ($f(\omega) / n$)

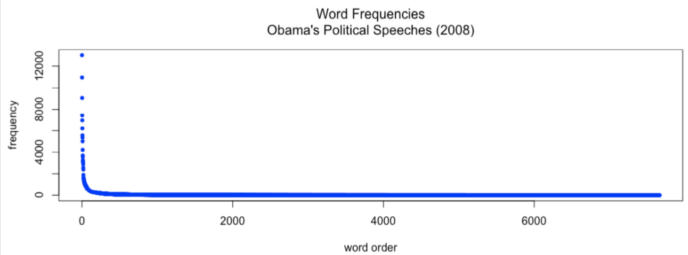
$$f(\omega) = \frac{c}{z^{\alpha}} = c \cdot z^{-\alpha}$$

- The value of *c* varies from one corpus to the next
- Based on Obama's Speeches (2008)
  max frequency: 13,027 ("the")
  number of types: 7,663
- Graph: from the most frequent ("the") to the less frequent

7

Zipf's Law

From Obama's speeches in 2008



8

Zipf's Law

- The Zipf's law could be more useful when considering the log-log relationship between the absolute frequency (f(ω)) and the rank (*z*) (For Zipf, α = 1)

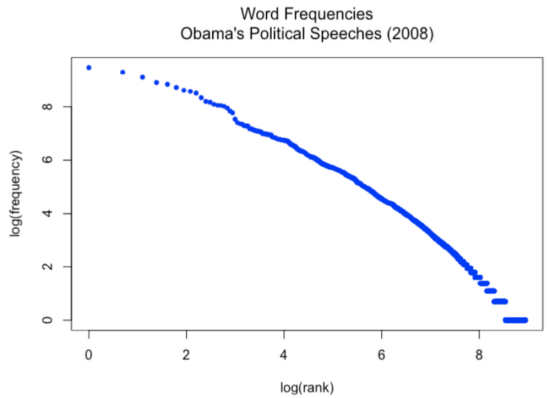
$$f(\omega) = \frac{c}{z^\alpha} = c \cdot z^{-\alpha}$$

we may obtain

$$log(f(\omega)) = log\left(\frac{c}{z^\alpha}\right)$$
$$= log(c) - \alpha \cdot log(z) = \beta - \alpha \cdot log(z)$$

- Zipf's law is an example of power law
  Another similar form is the 80-20 rule
- Property: scale invariant

Zipf's Law



Zipf's Law (French Language)

- From the French language
- Based on the newspaper *Le Monde* and ATS
- 34,508,866 tokens and 251,017 types (*vocables*)
- With the first 16 most frequent types, we cover around 30% of all French documents (news articles)

Rank	Word	Freq. f(ω)	Rel. Freq.	Cumul.	r x freq.
1	de	1,891,468	0.0548	0.0548	0.0548
2	la	1,062,987	0.0308	0.0856	0.0616
3	l	811,217	0.0235	0.1091	0.0705
4	le	807,145	0.0234	0.1325	0.0936
5	à	682,670	0.0198	0.1523	0.0989
6	les	657,241	0.0190	0.1713	0.1143
7	et	592,668	0.0172	0.1885	0.1202
8	des	584,412	0.0169	0.2054	0.1355
9	d	548,764	0.0159	0.2214	0.1431
10	en	477,379	0.0138	0.2352	0.1383
11	du	439,227	0.0127	0.2479	0.1400
12	a	409,561	0.0119	0.2598	0.1424
13	un	394,582	0.0114	0.2712	0.1486
14	une	335,561	0.0097	0.2809	0.1361
15	est	279,495	0.0081	0.2890	0.1215
16	dans	265,387	0.0077	0.2967	0.1231

Zipf's Law (German Language)

- Based on the newspaper *NZZ, Der Speigel,* and SDA
- 70,000,000 tokens and 1,081,681 types (*vocables*)
- With the first 16 most frequent types, we cover more than 20% of all German documents (news articles)
- The most frequent words are viewed as noisy from an information retrieval point of view
- But they correspond to style markers

13

Rank	Word	Freq.	Rel. Freq.	Cumul.	r x freq.
1	der	2,420,534	0.0346	0.0346	0.0346
2	die	2,407,558	0.0344	0.0690	0.0688
3	und	1,489,787	0.0213	0.0902	0.0639
4	in	1,243,042	0.0178	0.1080	0.0710
5	den	790,054	0.0129	0.1193	0.0564
6	von	668,300	0.0095	0.1288	0.0573
7	das	668,163	0.0095	0.1384	0.0668
8	mit	586,284	0.0084	0.1468	0.0670
9	im	568,533	0.0081	0.1549	0.0731
10	zu	556,061	0.0079	0.1628	0.0794
11	für	534,454	0.0076	0.1705	0.0840
12	des	489,420	0.0070	0.1775	0.0839
13	auf	481,672	0.0069	0.1843	0.0895
14	sich	456,291	0.0065	0.1909	0.0913
15	dem	429,675	0.0062	0.1970	0.0921
16	ein	421,569	0.0060	0.2030	0.0964

14

Zipf's Law

- On the other tail (the less frequent word types)
- Lot of word types with frequency = 1 (*hapax legomena*) and many with frequency = 2
- Number of word types: 7,663 (Obama), 7,792 (McCain)

Frequency	Obama'08		McCain'08	
1	2573	33.6%	2958	38.0%
2	1042	13.6%	1112	14.3%
3	556	7.3%	641	8.2%
4	446	5.8%	435	5.6%
5	308	4.0%	313	4.0%

15

Zipf's Law

- The Zipf' s law predict 50% *hapax legomena*
- Why?
  - Spelling errors (performance & diacritics)
  - Many proper names
  - but this is a general pattern
    few word types cover a large number of tokens
    large number of word types cover a few number of tokens
- Can we take a (large) sample of text and be sure to have all possible types?
- LNRE phenomenon: *Large Number of Rare Events*

16

Zipf's Law

- Example of *hapax legomena*

in McCain 2008	in Obama 2008
MI	AK
BMW	zionist
denial	WTO
bird	odd
richer	petrodollar
motel	Dupont
NALEO	Dehli

17

Question

Can we estimate the probability of occurrence of words? And sequences of them?

All words?

What are the benefits?

18

Language Model

- Estimating the occurrence probability of words

$$\sum_{x \in V^*} Prob[x] = 1 \quad and \ Prob[x] \geq 0 \quad \forall x \in V^*$$

- *Speech recognition* was the original motivation (Related problems are optical character recognition (OCR), handwriting recognition)
- The estimation techniques developed for this problem will be *very* useful for other problems in NLP (e.g. new model in IR).
- Difference between "A" and "a" or "The" and "the"?

19

Language Model

- How can we estimate the probability Prob[*s* = *This is a good deal*]?

- How can we estimate the underlying probabilities?
- How can we link the various words of the sentence?

Prob[*s*] = Prob[this | Δ] . Prob[is | Δ, this] ·
Prob[a | Δ, this, is] ·
Prob[good | Δ, this, is, a] ·
Prob[deal | Δ, this, is, a, good]

20

Language Model

- Using unigrams

$$Prob[w_i|w_1, w_2, \cdots, w_{i-1}] = Prob[w_i]$$

- Using bigrams (as approximations)

$$Prob[w_i \mid w_1, w_2, \cdots w_{i-1}] = Prob[w_i \mid w_{i-1}]$$

- Using trigrams (as approximations)

$$Prob[w_i \mid w_1, w_2, \cdots w_{i-1}] = Prob[w_i \mid w_{i-2}, w_{i-1}]$$

in our example, we obtained
Prob[*s*] = Prob[this | Δ] . Prob[is | Δ, this] . Prob[a | this, is] . Prob[good | is, a] . Prob[deal | a, good]

Language Model: Example

Unigram Model
Δ *This is a good deal* Δ
For unigram model (e.g., Prob[this] = 264 / 108,140 = 0.00244)

w_i	$C(w_i)$	$Prob[w_i]$
Δ	7,072	
this	264	0.00244
is	2,211	0.02045
a	2,482	0.02295
good	53	0.00049
deal	5	0.00005
Δ	7,072	

Language Model: Example

- Using the classical estimator for bigrams
  $C(w_k)$ = count / frequency of word w_k

$$Prob[w_i|w_{i-1}] = \frac{C(w_{i-1}, w_i)}{\sum_w C(w_{i-1}, w)} = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

Δ *This is a good deal* Δ

Language Model: Example

Δ *This is a good deal* Δ
For bigram model (e.g., Prob[this| Δ] = 0.0188 = 133 / 7072)

Unigrams Bigrams

w_i	$C(w_i)$	$Prob[w_i]$	w_{i-1}, w_i	$C(w_{i-1}, w_i)$	$Prob[w_i \mid w_{i-1}]$
Δ	7,072		Δ this	133	0.0188
this	264	0.00244	this is	14	0.0530
is	2,211	0.02045	is a	24	0.0109
a	2,482	0.02295	a good	2	0.0008
good	53	0.00049	good deal	0	0
deal	5	0.00005	deal Δ	1	0.2
Δ	7,072				

Do we have a perfect solution?

Sparse Data Problem

- We have a lot of counts = 0
  and thus many estimations = 0
- Data sparseness is a serious and common problem in statistical NLP.
- The probability of a sequence is zero if it contains unseen elements (types, bigram)
- Problem 1: Zero counts
  If *n*-gram ω_y does not occur in the training set, does that mean that it should have probability zero?
- Problem 2: Low frequency *n*-grams
  if *n*-gram ω_x occurs twice and *n*-gram ω_y occurs once, is ω_x really twice as likely as ω_y?

25

Smoothing techniques

This is a black art in Natural Language Processing (NLP)

26

Smoothing the Estimates

- We have in the corpus $\{\omega_x\omega_a, \ldots \omega_x\omega_b, \ldots \omega_x\omega_b\}$
- Should we conclude
  - Prob[ω_a | ω_x] = 1/3? *reduce this*
  - Prob[ω_b | ω_x] = 2/3? *reduce this*
  - Prob[ω_c | ω_x] = 0/3? *increase this*
- *Discount* the positive counts somewhat
- *Reallocate* that probability to the zeroes
- Especially if the *denominator* is small …
  - 1/3 probably too high, 100/300 probably about right
- Especially if *numerator* is small …
  - 1/300 probably too high

27

Language Model: Example

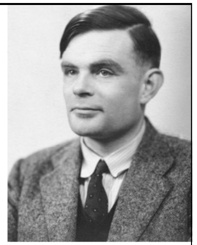
Laplace's rule

$$Prob[w_i|w_{i-1}] = \frac{C(w_{i-1}, w_i) + 1}{\sum_w C(w_{i-1}, w) + 1} = \frac{C(w_{i-1}, w_i) + 1}{C(w_{i-1}) + |V|}$$

w_{i-1}, w_i	$C(w_{i-1}, w_i)$	$C(w_{i-1})$ + |V|	Prob[w_i | w_{i-1}]
Δ this	133 + 1	7,072 + 8,635	0.0085
this is	14 + 1	264 + 8,635	0.0017
is a	24 + 1	2211 + 8,635	0.0023
a good	2 + 1	2482 + 8,635	0.0003
good deal	0 + 1	53 + 8,635	0.0001
deal Δ	1 + 1	5 + 8,635	0.0002

28

Good-Turing Smoothing



- Intuition: Can judge rate of novel events by rate of singletons
- If we have seen a lot of singletons, then new novel events are also likely.
- Here we present the simplest Good-Turing scheme
  More complex models do exist!
- Let N_c = the number of *n*-grams that occurred exactly *c* times in the corpus.
  - e.g., N_0 = number of unseen *n*-grams
  - e.g., N_1 = number of *n*-grams seen once
- Let $N = \sum_c N_c$ total # of training tokens

Good-Turing Smoothing

- The frequency of *n*-grams occurring *c* times is re-estimated as: $c^* = (c+1) \cdot \frac{N_{c+1}}{N_c}$
- Unseen *n*-grams is: $c^* = \frac{N_1}{N_0}$

  and the *n*-grams seen once: $c^* = \frac{2 \cdot N_2}{N_1}$

  and the total number of bigrams = $|V|^2$

Good-Turing Smoothing

- *Nineteen eighty-four* contain 37,365 unique bigrams and 5,820 bigrams seen twice…
  Its vocabulary of 8,635 words generates $8{,}635^{\wedge 2}$ = 74,563,225 bigrams whose 74,513,701 are unseen.

- Unseen bigram: (37,365 / 74,513,701) = 0.0005
  and unique bigrams: (2 · 5,820 / 37,365) = 0.31

Good-Turing Smoothing

Reestimate only if $N_c < 10$

w_i, w_{i+1}	$C(w_i,w_{i+1})$	$c^*(w_i,w_{i+1})$	$P[w_{i+1} \mid w_i]$
Δ this	133	133	
this is	14	14	
is a	24	24	
a good	2	→ *1.09*	
good deal	0	→ *0.0005*	
deal Δ	1	→ *0.31*	

Good-Turing Smoothing

$$Prob[\Delta|this] = \frac{C(\Delta, this)}{C(this)} = \frac{133}{7,072} = 0.0188$$

w_i, w_{i+1}	$C(w_i,w_{i+1})$	$c^*(w_i,w_{i+1})$	$P[w_{i+1} \| w_i]$
Δ this	133	133	133/7,072 = 0.0188
this is	14	14	14/264 = 0.0530
is a	24	24	24/2,211 = 0.0109
a good	2	→ *1.09*	1.09 / 2,482 = 0.0004
good deal	0	→ *0.0005*	0.0005 / 53 = 0.00001
deal Δ	1	→ *0.31*	0.31 / 5 = 0.062

Language Model: *Like*

- Another look at the language model
- The verb *like*
  Appears 97,179 with a nominal subject
  and 52,904 with a direct object.

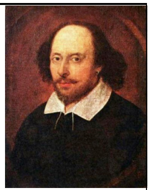
  As subject:

I	50%
you	14%
they	4%
we	4%
people	2%

  As object:

it	12%
what	4%
idea	2%
they	2%

Applications



Authorship Attribution (AA): Who wrote this text?

We have a set of documents written by $A_1, A_2, \ldots, A_n$.

We have a disputed text Q. Who is the author of this text?

Solution: Compute a distance between the different possible authors

Possible contexts:

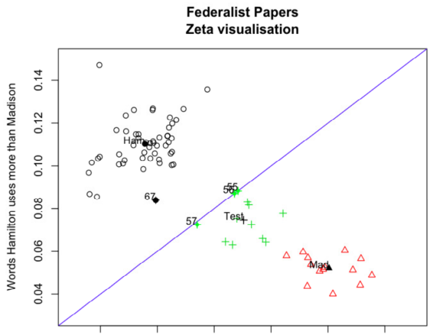
1. closed-set: The true author is in the list
2. open-set: The true author might be in the list or it is another unknown
3. profiling: Infer some socio-demographic information about the author

Federalist Papers

	Hamilton		Madison	
Rank	**Word**	**Freq.**	**Word**	**Freq**
1	the	10,293	the	3,907
2	,	7,483	,	2,805
3	of	7,149	of	2,318
4	to	4,495	to	1,253
5	.	2,929	and	1,168
6	in	2,778	.	1,039
7	and	2,681	in	808
8	a	2,476	a	771
9	be	2,270	be	755
10	that	1,679	that	542

Who is the Author?

Percentage of words used mainly by Hamilton



Percentage of words used mainly by Madison

Federalist Papers

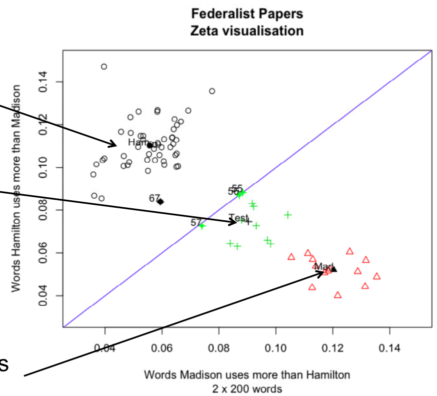
index	Hamilton	index	Madison
1.76	upon	1.42	existing
1.60	kind	1.40	fully
1.60	community	1.38	clearly
1.45	matter	1.37	among
1.44	easy	1.37	according
1.44	execution	1.36	indefinite
1.42	intended	1.36	consequently
1.42	done	1.36	whilst
1.42	sometimes	1.35	confederation
1.40	circumstances	1.34	absolutely

Who is the Author?

Hamilton's articles

The 12 disputed articles

Madison's articles



Profiling: Gender & Age

Female vs. Male?
Teen , Twenties, or Thirties?

```
Yesterday we had our second jazz
competition. Thank God we weren't
competing. We were sooo bad. Like, I
was so ashamed, I didn't even want to
talk to anyone after. I felt so
rotton, and I wanted to cry,
but...it’s ok.
```

40

Profiling: Gender & Age

Female vs. Male?
Teen , Twenties, or Thirties?

```
My gracious boss had agreed to let me
have one week off of "work." He did
finally give me my report back after
eight freakin' days! Now I only have
the rest of this week and then one
full week after my vacation to finish
this damned thing.
```

41

Profiling

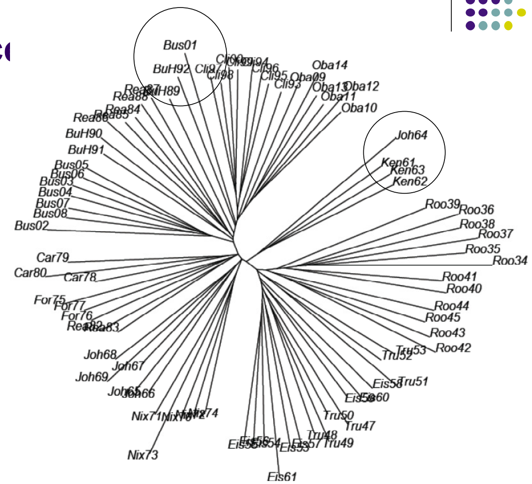
	Male	Female
job	68.1±0.6	56.5±0.5
money	43.6±0.4	37.1±0.4
sports	31.2±0.4	20.4±0.2
tv	21.1±0.3	15.9±0.2
sex	32.4±0.4	43.2±0.5
family	27.5±0.3	40.6±0.4
eating	23.9±0.3	30.4±0.3
friends	20.5±0.2	25.9±0.3
sleep	18.4±0.2	23.5±0.2
pos-emotions	248.2±1.9	265.1±1.2
neg-emotions	159.5±1.3	178±1.4

42

State of the Union Addresses

- Governmental speeches (≠ electoral)
  - 81 addresses (annual)
  - 13 US presidents
- For the Congress & nation
  - State of the Union / world
  - Legislative propositions
- Questions
  - Can we assign each speech to his presidency?
  - What is specific to Obama?

Distance



Characteristic Sentences

Which US president wrote …

"The *American* people deserve a tax code that helps small businesses spend less time filling out complicated forms, and more time expanding and hiring; a tax code that ensures billionaires with high-powered accountants can not pay a lower rate than their hard-working secretaries; a tax code that lowers incentives to move *jobs* overseas, and lowers tax rates for businesses and manufacturers that create *jobs* right here in America".

45

Characteristic Sentences

"Our own *objectives* are clear; the *objective* of smashing the militarism imposed by *war* lords upon their enslaved peoples, the *objective* of liberating the subjugated Nations, the *objective* of establishing and securing freedom of speech, freedom of religion, freedom from want, and freedom from fear everywhere in the world".

46

Language Models

J. Savoy
Université de Neuchâtel

C. D. Manning & H. Schütze: *Foundations of statistical natural language processing*. The MIT Press. Cambridge (MA)

P. M. Nugues: *An introduction to language processing with Perl and Prolog*. Springer. Berlin

R. H. Baayen : *Word Frequency Distributions*. Kluwer. Drodrecht

47