OA Inc

9/15/16 CS 585 UMass Amherst

(1) Evaluating LM

(2) Pseudocount Smoothing :

(3) Noam Chomsky is still unhappy

Eval

(1) Main: High test prob

(2) PPL def

(3) Shannon Game

(4) PPL interp

How to evaluate your LM?

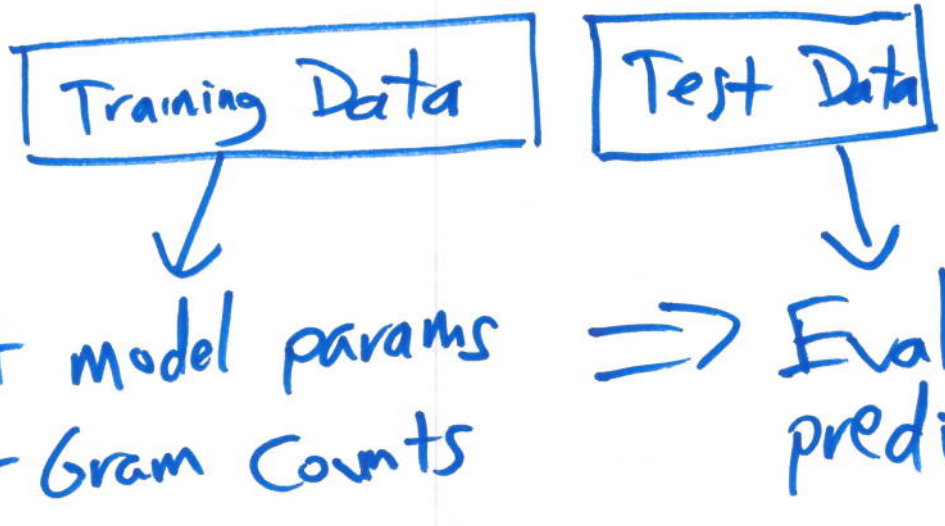
Intrinsic vs **Extrinsic**

Intrinsic → Eval LM on its own

Easier → Check LM's probabilistic predictions on new **test data**

Extrinsic → Plug in to a system e.g. machine translation

Best



Want: high prob of new text data

$w_1, w_2, \ldots w_N$

Large $P(w_1, w_2 \ldots w_N) \Leftrightarrow$ Large $\log P(w_1 \ldots w_N)$

$$\log P(w_1, w_2, w_3 \ldots w_N)$$

$$= \log \prod_{i=1}^{N} P(w_i \mid w_1 \ldots w_{i-1})$$

$$= \sum_i \log P(w_i \mid w_1 \ldots w_{i-1})$$

$$\text{Perplexity} = e^{\underbrace{-\frac{1}{N} \log P(w_1 \ldots w_N)}_{\text{Neg. Avg. Log Prob per Token}}}$$

Low PPL $\Leftrightarrow$ High Prob

[Shannon Game & spreadsheet here]

Book Def. PPL $= \sqrt[N]{\frac{1}{P(w_1 \ldots w_N)}}$

$= [P(w_1 \ldots w_N)]^{-1/N}$

Claim: equivalent to $\exp(\frac{-1}{N} \log P(\ldots))$

Interp: PPL as branching factor

PPL of uniform dist ? $\Rightarrow$ PPL $= |V|$

$|V| = 4$ Pred: $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$

$PPL = e^{-\frac{1}{N} \log P(x)}$

$= e^{-\frac{1}{4} \log \frac{1}{4}} = e^{\log 4} = 4$

PPL measures uncertainty

equiv. ~~uniform~~ # sides of uniform die

Lower perplexity = better model

- Training 38 million words, test 1.5 million words, WSJ

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Zeros

Problem!

→ Very natural in language

Sparsity of language

- Training set:
  - … denied the allegations
  - … denied the reports
  - … denied the claims
  - … denied the request

- Test set
  - … denied the offer
  - … denied the loan

P("offer" | denied the) = 0

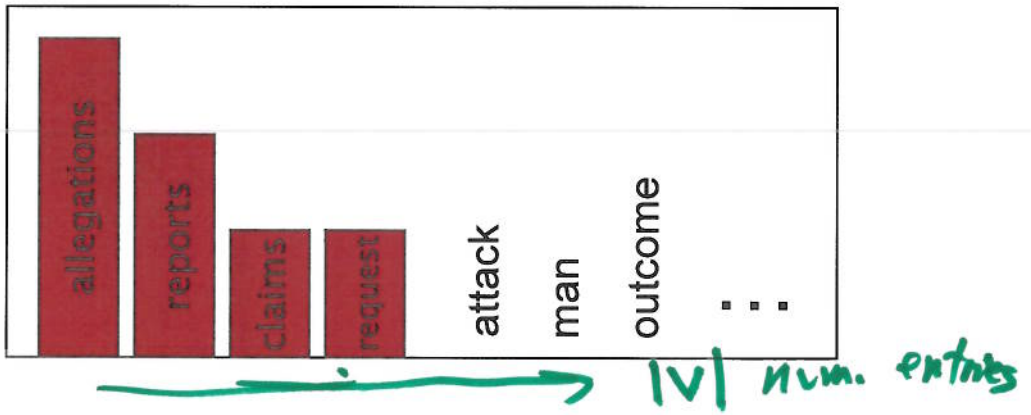
$P(\text{test sent}) = P(w_1)\,P(w_2|w_1)\,\underbrace{P(w_3|w_1 w_2)}_{=0} \cdots$

$=0$

The intuition of smoothing (from Dan Klein)

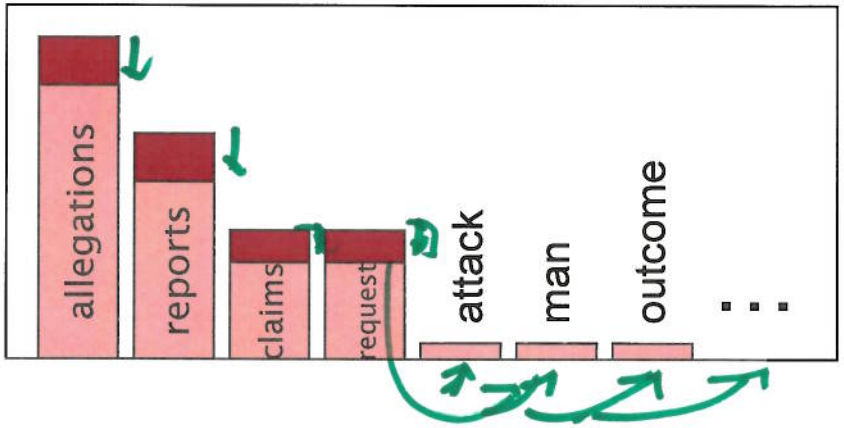
- When we have sparse statistics:

  P(w | denied the)
  3 allegations
  2 reports
  1 claims
  1 request
  7 total



- Steal probability mass to generalize better

  P(w | denied the)
  2.5 allegations
  1.5 reports
  0.5 claims
  0.5 request
  2 other
  7 total



out-of-vocab words???

OOV

Add-one estimation

- Also called Laplace smoothing
- Pretend we saw each word one more time than we did
- Just add one to all the counts!
- MLE estimate:

$$P_{MLE}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

- Add-1 estimate:

$$P_{Add-1}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)+1}{c(w_{i-1})+|V|}$$

Generally Add-α smoothing

$\alpha = 0.1$ $\alpha = 0.01$

$\alpha = 100$

Berkeley Restaurant Corpus: Laplace smoothed bigram counts

	i	want	to	eat	chinese	food	lunch	spend
i	6	828	1	10	1	1	1	3
want	3	1	609	2	7	7	6	2
to	3	1	5	687	3	1	7	212
eat	1	1	3	1	17	3	43	1
chinese	2	1	1	1	1	83	2	1
food	16	1	16	1	2	5	1	1
lunch	3	1	1	1	1	2	1	1
spend	2	1	2	1	1	1	1	1

$$\frac{C(w_{n-2}\, w_{n-1}\, w_n) + 1}{C(w_{n-1}\, w_n) + V}$$

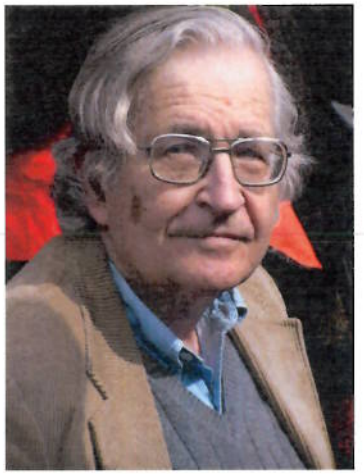
Laplace-smoothed bigrams

$$P^*(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n) + 1}{C(w_{n-1}) + V}$$

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Challenging example

Noam Chomsky (*Syntactic Structures*, 1957):



Sentences (1) and (2) are equally nonsensical, but any speaker of English will recognize that only the former is grammatical.

(1) Colorless green ideas sleep furiously.
(2) Furiously sleep ideas green colorless.

[T]he notion "grammatical in English" cannot be identified in any way with the notion "high order of statistical approximation to English". It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English.