Linguistics 696f
Hammond
Spring '03

Smoothing

A. Overview

(1)
a. Sparse data
b. "Add-one" smoothing
c. Witten-Bell smoothing
d. Deleted interpolation
e. Good-Turing smoothing
f. Held-out Estimation
g. Software

B. Sparse data

(2)

story	author	tokens	types	hapax
The Adventure of the Bruce-Partington Plans	Sir Arthur Conan Doyle	11096	2106	1184
The Disappearance of Lady Frances Carfax	Sir Arthur Conan Doyle	7947	1744	1028
The Call of the Wild	Jack London	32462	4731	2491
White Fang	Jack London	73992	6556	2938

(3)

	Plans	Carfax	Wild	Fang
Plans	0	891	3605	5240
Carfax	1253	0	3770	5438
Wild	980	783	0	3590
Fang	790	626	1765	0

C. "Add-one" smoothing

(4) maximum likelihood: $\mathrm{p}(w_i) = \frac{|w_i|}{\sum |w|}$

(5) $\mathrm{p}^*(w_i) = \frac{|w_i|+1}{N+V}$

D. Witten-Bell smoothing

(6) Reserved mass: $\frac{T}{T+N}$, where T is the number of n-gram types and N is the total number of n-gram tokens.

(7) Discounting: $\mathrm{p}^*(w_i) = \frac{|w_i|}{N+T}$ if $|w_i| > 0$

(8) $c_i^* = \begin{cases} \frac{T}{Z} \times \frac{N}{N+T} & \text{if } c_i{=}0 \\ c_i \times \frac{N}{N+T} & \text{if } c_i{>}0 \end{cases}$, where Z is the number of n-grams with a count of 0.

E. Deleted interpolation

(9) $\hat{\mathrm{p}}(w_j|w_i) = \lambda_1 \mathrm{p}(w_j) + \lambda_2 \mathrm{p}(w_j|w_i)$ where $\lambda_1 + \lambda_2 = 1$

F. Good-Turing smoothing

(10) $c^* = (c+1)\frac{N_{c+1}}{N_c}$

(11) Katz backoff: $c^* = \frac{(c+1)\frac{N_{c+1}}{N_c} - c\frac{(k+1)N_{k+1}}{N_1}}{1 - \frac{(k+1)N_{k+1}}{N_1}}$, for $1 \leq c \leq k$.

G. Held-out Estimation

(12) N_r: the number of n-grams with frequency r in the training text.
C_2: the count for this n-gram in the held-out text

(13) $T_r = \sum C_2(w_1 \cdots w_n)$

(14) $\mathrm{p_{ho}}(w_1 \cdots w_n) = \frac{T_r}{N_r N}$ where $C(w_1 \cdots w_n) = r$

H. Software

(15) All software is invoked on the command-line as follows: `perl program-name`
If arguments are required, a suitable error message is displayed.

(16)
a. `hapax.pl`: counts the hapax legomena in a text.
b. `novelwords.pl`: counts novel words in a text.
c. `addonecross.pl`: calculates the cross-entropy of a text using add-one smoothing.
d. `addx.pl`: calculates the cross-entropy of a text using "add-x" smoothing.

References

Charniak, Eugene. 1993. *Statistical Language Learning*. Cambridge: MIT Press.

Chen, S. F., & J. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University.

Katz, Slava. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing* 35.400–401.