Linguistics 696f
Hammond
Spring '03

Smoothing continued

A. Overview

(1) a. Good-Turing smoothing
b. Witten-Bell smoothing

B. Good-Turing smoothing

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

(3) 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 $1 \leq c \leq k$.

(4) A hypothetical case:
70 words, 4900 possible bigrams, 2556 bigrams in text, 1661 distinct occurring bigrams

(5)

bigrams	counts	p(w_1w_2)	c^*	p*(w_1w_2)
1	6	.00234	?	?
10	5	.00195	.6	.000234
50	4	.00156	1	.000391
100	3	.00117	2	.000782
500	2	.000782	.6	.000234
1000	1	.000391	1	.000391
3239	0	0	.308	.00012

(6) $c_0^* = (c_0+1)\frac{N_{c_0+1}}{N_{c_0}} = 1 \times \frac{1000}{3239} = .308$

(7) $\text{p}^*(w_1w_2) = \frac{c_0^*}{N} = \frac{.308}{2556} = .00012$, if $|w_1w_2| = 0$

(8) Note that the discounted value for the highest case can be set on the assumption that these values exhibit a probability distribution, e.g. $1 - (.000234 \times 10) - (.000391 \times 50) - (.000782 \times 100) - (.000234 \times 500) - (.000391 \times 1000) - (.00012 \times 3239) =$?

C. Witten-Bell smoothing

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

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

(11) Reserved mass: $\frac{1661}{1661+2556} = .393$, divided evenly among nonoccurring n-grams.

(12)

bigrams	counts	$\mathrm{p}(w_1w_2)$	$\mathrm{p}^*(w_1w_2)$
1	6	.00234	.00142
10	5	.00195	.00118
50	4	.00156	.000948
100	3	.00117	.000711
500	2	.000782	.000472
1000	1	.000391	.000237
3239	0	0	.000121

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.