Spellcheck-2

Train and Test

- Training on: Brown corpus (1 million words)
- Testing on: Wall Street Journal corpus (3/4 million words)

Baseline

$P(c_{-k} \ldots c_{-1}, c_1 \ldots c_k \mid w) \cdot P(w)$

Confusion set	No. of training cases	No. of test cases	Most frequent word	Baseline
whether, weather	331	245	whether	0.922
I, me	6125	840	I	0.886
its, it's	1951	3575	its	0.863
past, passed	385	397	past	0.861
than, then	2949	1659	than	0.807
being, begin	727	449	being	0.780
effect, affect	228	162	effect	0.741
your, you're	1047	212	your	0.726
number, amount	588	429	number	0.627
council, counsel	82	83	council	0.614
rise, raise	139	301	rise	0.575
between, among	1003	730	between	0.538
led, lead	226	219	led	0.530
except, accept	232	95	except	0.442
peace, piece	310	61	peace	0.393
there, their, they're	5026	2187	there	0.306
principle, principal	184	69	principle	0.290
sight, site, cite	149	44	sight	0.114

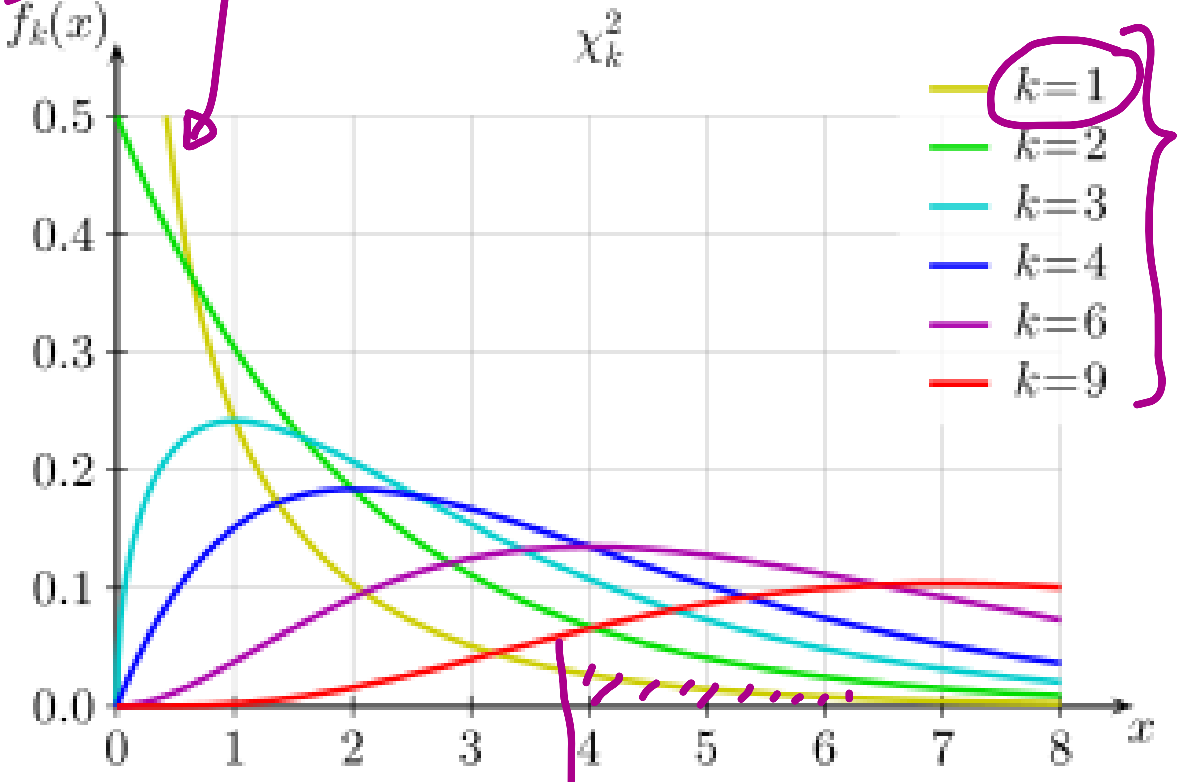
Chi-square distribution

χ^2

χ^2 4

$\sum_{i=1} \frac{(O_i - E_i)^2}{E_i}$

95%



θ

Is the value (4) more extreme than the critical value

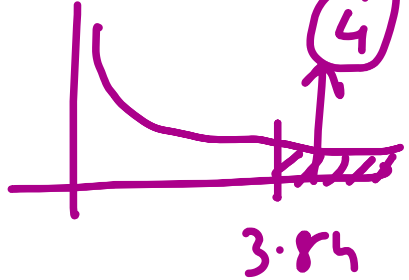
$\chi^2_c \rightarrow \theta$

α

$\int_{\theta} y \, dx = 0.05$

Chi-squared (Wikipedia)

4



α

$\int_{3.84} \chi^2_1 \, dx = 0.05$

Degrees of freedom (df)	χ^2 value [13]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
	Non-significant								Significant		

icing

desert 0
dessert 3

The algorithm

Training phase

(1) Propose all words as candidate context words.
(2) Count occurrences of each candidate context word in the training corpus.
(3) Prune context words that have insufficient data or are uninformative discriminators.
(4) Store the remaining context words (and their associated statistics) for use at run time.

Run time

(1) Initialize the probability for each word in the confusion set to its prior probability.
(2) Go through the list of context words that was saved during training. For each context word that appears in the context of the ambiguous target word, update the probabilities.
(3) Choose the word in the confusion set with the highest probability.

Confusion set	Baseline	Cwords ±3	Cwords ±6	Cwords ±12	Cwords ±24
whether	0.922	0.902	0.922	0.927	0.922
I	0.886	0.914	0.893	0.883	0.851
its	0.863	0.862	0.795	0.743	0.702
past	0.861	0.861	0.849	0.801	0.743
than	0.807	0.931	0.901	0.896	0.855
being	0.780	0.791	0.795	0.793	0.755
effect	0.741	0.747	0.741	0.759	0.716
your	0.726	0.816	0.783	0.774	0.736
number	0.627	0.646	0.622	0.636	0.639
council	0.614	0.639	0.614	0.602	0.614
rise	0.575	0.575	0.575	0.585	0.498
between	0.538	0.759	0.697	0.671	0.586
led	0.530	0.530	0.530	0.521	0.557
except	0.442	0.695	0.526	0.516	0.558
peace	0.393	0.754	0.705	0.574	0.574
there	0.306	0.726	0.623	0.557	0.466
principle	0.290	0.290	0.290	0.290	0.435
sight	0.114	0.455	0.250	0.364	0.318
Avg no. of context words		27.9	36.9	55.9	92.9

Confusion set	Baseline	Collocs ≤ 1	Collocs ≤ 2	Collocs ≤ 3
whether	0.922	0.939	0.931	0.931
I	0.886	0.979	0.981	0.980
its	0.863	0.943	0.945	0.950
past	0.861	0.919	0.909	0.909
than	0.807	0.966	0.965	0.966
being	0.780	0.853	0.853	0.842
effect	0.741	0.821	0.821	0.821
your	0.726	0.877	0.887	0.887
number	0.627	0.646	0.646	0.681
council	0.614	0.663	0.639	0.639
rise	0.575	0.807	0.807	0.807
between	0.538	0.699	0.730	0.733
led	0.530	0.849	0.840	0.863
except	0.442	0.800	0.789	0.789
peace	0.393	0.869	0.869	0.852
there	0.306	0.911	0.932	0.932
principle	0.290	0.841	0.812	0.812
sight	0.114	0.341	0.318	0.318
Avg no. of collocations		33.9	263.1	985.4

Context word	*peace*	*piece*
corps	49	1
peace	41	1
united	20	0
nations	15	0
our	27	1
heart	12	0
justice	12	0
state	12	0
american	11	0
aid	11	0
international	11	0
women	10	0
war	20	1
world	40	3
piece	1	15
over	1	14
must	11	1
great	11	1
under	10	1
how	10	1
⋮		
two	5	12
for	83	38
about	4	9
every	4	9
little	5	10
long	6	11
one	14	23
the	179	113
so	9	14
;	16	22
Total occurrences	184	126

Collocation	*peace*	*piece*
__ corps	47	0
DET __ corps	32	0
ADV __ corps	28	0
the __ corps	27	0
__ and	22	0
__ of NS	2	60
the __ NS	37	1
a __ PREP	1	35
PREP __ of	1	34
a __ of	1	34
for __	16	0
__ and NS	16	0
DET __ NP	32	1
NS __ of	2	45
__ corps NS	14	0
PREP __ CONJ	14	0
the __ NP	27	1
V CONJ __	13	0
__ NS PUNC	13	0
__ of V	1	25
⋮		
CONJ ADJ __	4	9
the NS __	4	9
NS ADJ __	13	26
ADV NS __	12	23
PREP NS __	17	31
ADV __ PREP	12	22
ADJ ADJ __	9	14
NS __	62	79
ADJ __	46	54
NS NS __	29	32
Total occurrences	184	126

Russel's Soundex

The alphabet was phonetically divided into categories:

Oral resonants A, E, I, O, U, Y.
Labials and labio-dentals B, F, P, V.
Gutterals and sibilants C, G, K, Q, S, X, Z .
Dental-mutes D, T/
Palatal-fricative L.
Labio-nasal M.
Den to or lingua-nasal N.
Dental fricative R.

Russel also described a few additional rules to complete the indexing:

- The initial letter of the word is always kept.
- Two consecutive letters that had the same code are considered as a single letter (e.g. "BB" is the same as just "B")
- If a word ended with "GH", "S" or "Z' those letters were discarded.
- Only the first occurrence of a vowel (Group 1) is counted.

Ack:http://www.datamanagementgroup.com/Resources/Articles/Article_IntroductionToDoubleMetaphone.asp

Soundex Revised

- U.S. Government Soundex Table
  1. B,F,P,V
  2. C,G,J,K,Q,S,X,Z
  3. D, T
  4. L
  5. M, N
  6. R
- Examples
  - Johnson = J525
  - Miller = M460
  - Ricardo = R263
  - Peters = P362

Jensen

JNSN

$\boxed{J525}$

Ack:http://www.datamanagementgroup.com/Resources/Articles/Article_IntroductionToDoubleMetaphone.asp

Metaphone

- Find out from:
  - http://www.lanw.com/java/phonetic/default.htm

Selecting candidates for correction

The n-gram approach using inverted files

The next idea in spellcheck: Web n-grams

language modeling

ceramics collection and 43
ceramics collection at 52
ceramics collection is 68
ceramics collection of 76
ceramics collection | 59
ceramics collections , 66
ceramics collections . 60
ceramics combined with 46
ceramics come from 69
ceramics comes from 660
ceramics community , 109
ceramics community . 212
ceramics community for 61
ceramics companies . 53
ceramics companies consultants 173

Web as corpus

Two important papers

- Shane Bergsma, Dekang Lin, and Randy Goebel. 2009. *Web-scale n-gram models for lexical disambiguation*. In IJCAI.

- W. Xu, J. Tetreault, M. Chodorow, R. Grishman, and L. Zhao. 2011. Exploiting syntactic and distributional information for spelling correction with webscale n-gram models. In EMNLP.