de Marcken's version of EM

John A Goldsmith

Expected counts (soft counts)

Let's calculate the soft counts in a particular string of the words that happen to be in our lexicon. ("happen" here means that we will talk later about deciding which words should be there.)

A distribution

Counts summed to 1,000,000.

word	count	frequency	plog
A	15 600	.0156	6
B	15 600	.0156	6
C	15 600	.0156	6
...			
HE	62 500	0.0625	4
HER	62 500	0.0625	4
THE	125 000	0.125	3
HERE	31 125	0.031 25	5
THERE	31 125	0.03125	5
RENT	7 810	0.00781	7
IS	62 500	0.0625	4
TIS	3 906	0.003906	8
DUE	7 810	0.00781	7

Compute α (alpha)

t	term 1	prob	term 2	prob	partial	total
1		1				
2		1	T	0.015 6	0.015 6	0.015 6
3	T	0.015 6	H	0.015 6	.000 244	
3	TH					.000 244
4		1	THE	0.125	0.125	
4	T	0.015 6	HE	0.062 5	0.000 975	
4	TH	.000 244	E	0.015 6	3.80×10^{-5}	
4	THE					.125 978
5	T	0.015	HER	0.062 5	0.000 937	
5	TH		ER		0.000	
5	THE	.125 978	R	.015 6	0.001 96	
5	THER					0.002 902

Compute α (alpha)

t	term 1	prob	term 2	prob	partial	total
5	T	0.015	HER	0.062 5	0.000 937	
5	TH		ER		0.000	
5	THE	.125 978	R	.015 6	0.001 96	
5	THER					0.002 902
6		1.0	THERE	1	0.312 5	
6	T	0.015 6	HERE	0.031 25	0.000 975	
6	TH		ERE			
6	THE		RE			
6	THER	0.002 90	E	0.015 6	4.52×10^{-4}	
6	THERE					0.313

Compute α (alpha)

W t	term 1	prob	term 2	prob	partial
7	THERE	0.313	N	0.015 6	0.004 89
7	THEREN		**total:**	0.004 89	
8	THE	.002 9	RENT	.0078 1	.000 015
W 8	THEREN	.004 89	T	.015 6	.000 076
8	THERENT		**total:**	0.000 091	
9	THERENT	.000 091	I	0.015 6	.000 001 4
9	THERENTI		**total:**	0.000 001 4	
10	THERENT	.000 091	IS	0.062 5	.000005688
10	THERENTI	.000 001 4	S	0.015 6	.000 000 022
10	THERENTIS		**total:**	.00000571	
11	THERENTIS		D	.000 000 089	
11	THERENTISD		**total:**	.000 000 089	

Compute α (alpha)

t	term 1	prob	term 2	prob	partial
11	THERENTIS		D	0.015 6	.000 000 089
11	THERENTISD		**total**:	.000 000 089	
12	THERENTISD	.000 000 089	U	0.015 6	1.38 x 10^{-}
12	THERENTISDU		**total**:	1.38 x 10^{-9}	
13	THERENTIS	.00000571	DUE	0.00781	4.45 x 10^{-}
13	THERENTISDU	1.38 x 10^{-9}	E	0.0156	2.15 * 10^{-}
13	THERENTISDUE		**total**:	4.46 x 10^{-8}	

Soft count of *due*

$\text{alpha}(\text{therentis}) = 0.000\ 005\ 71$

$\text{beta}(\text{due}) = 0.007\ 81$

$\frac{4.46*10^{-8}}{4.46x10^{-8}}$

Soft count of *is*

alpha(therent) = 0.000 091
beta(due) = 0.007 81 (sum of 0.007 81 and 0.000 000 38)
pr(is) = 0.062 5
$0.000091 \times 0.0625 \times 0.00781 = 4.44 \times 10^{-8}$
soft count of *is* = $\frac{4.44 \times 10^{-8}}{4.46 * 10^{-8}}$
$4.46 * 10^{-8}$
$\frac{4.46 * 10^{-8}}{4.46 x 10^{-8}}$ =0.995 51

Viterbi parse is different

t	term 1	prob	term 2	prob	partial
1		1			
2		1	T	0.015 6	0.015 6
2	Best parse is just T				
3	T	0.015 6	H	0.015 6	.000 244
3	Best parse is T-H				
4		1	THE	0.125	0.125
4	T	0.015 6	HE	0.062 5	0.000 975
4	TH	.000 244	E	0.015 6	3.80 x 10^{-5}
4	Best parse is THE				
5	T	0.015	HER	0.062 5	0.000 937
5	TH		ER		0.000
5	THE	.125 978	R	.015 6	0.001 96
5	Best parse is T-HER				

Viterbi 2

t	term 1	prob	term 2	prob	partial
5	T	0.015	HER	0.062 5	0.000 937
5	TH		ER		0.000
5	THE	.125 978	R	.015 6	0.001 96
5	Best parse is T-HER				
6		1.0	THERE	1	0.312 5
6	T	0.015 6	HERE	0.031 25	0.000 975
6	TH		ERE		
6	THE		RE		
6	THER	0.002 90	E	0.015 6	$4.52x10^{-}$
6	Best parse is THERE				

Viterbi 3

t	term 1	prob	term 2	prob
7	THERE	0.313	N	0.015 6
7	Best parse is THERE-N			
8	THE	.002 9	RENT	.0078 1
8	THEREN	.004 89	T	.015 6
8	Best parse is THE-RENT			
9	THERENT	.000 091	I	0.015 6
9	Best parse is [THE-RENT]-I			
10	THERENT	.000 091	IS	0.062 5
10	THERENTI	.000 001 4	S	0.015 6
10	Best parse is [THE-RENT]-IS			
11	THERENTIS		D	.000 000 089
11	THERENTISD			.000 000 089
11	Best parse is [THE-RENT-IS]-D			

Viterbi 4

t	term 1	prob	term 2	prob
11	THERENTIS		D	0.015 6
11	Best parse is [THE-RENT-IS]-D			
12	THERENTISD	.000 000 089	U	0.015 6
12	Best parse is [THE-RENT-IS-D]-U			
13	THERENTIS	.00000571	DUE	0.00781
13	THERENTISDU	1.38×10^{-9}	E	0.0156
13	Best parse is [THE-RENT-IS]-DUE			