Reducing perplexities

(and word error rates)
with interpolation factors

CLS | Centre for Language Studies
Radboud University

Reducing perplexities and word error rates by improving the interpolation factors of a Bayesian skipgram language model

Louis Onrust
Centre for Language Studies, Radboud University
Center for Processing Speech and Images, KU Leuven

l.onrust@let.ru.nl
github.com/naiaden

Language Model Recap

HPYPLMs

- first pass decoder
- bayesian language model
- hpyplm follows the analogy of the chinese restaurant process - srilm follows the analogy of a soup kitchen
- effect domains - report perplexities - report word error rates

Backoff strategies

- ngram
- full
- limited

Interpolation factors

- uniform - npref - value
- count
- perplexity - entropy - mle
- random

Perplexities on English data

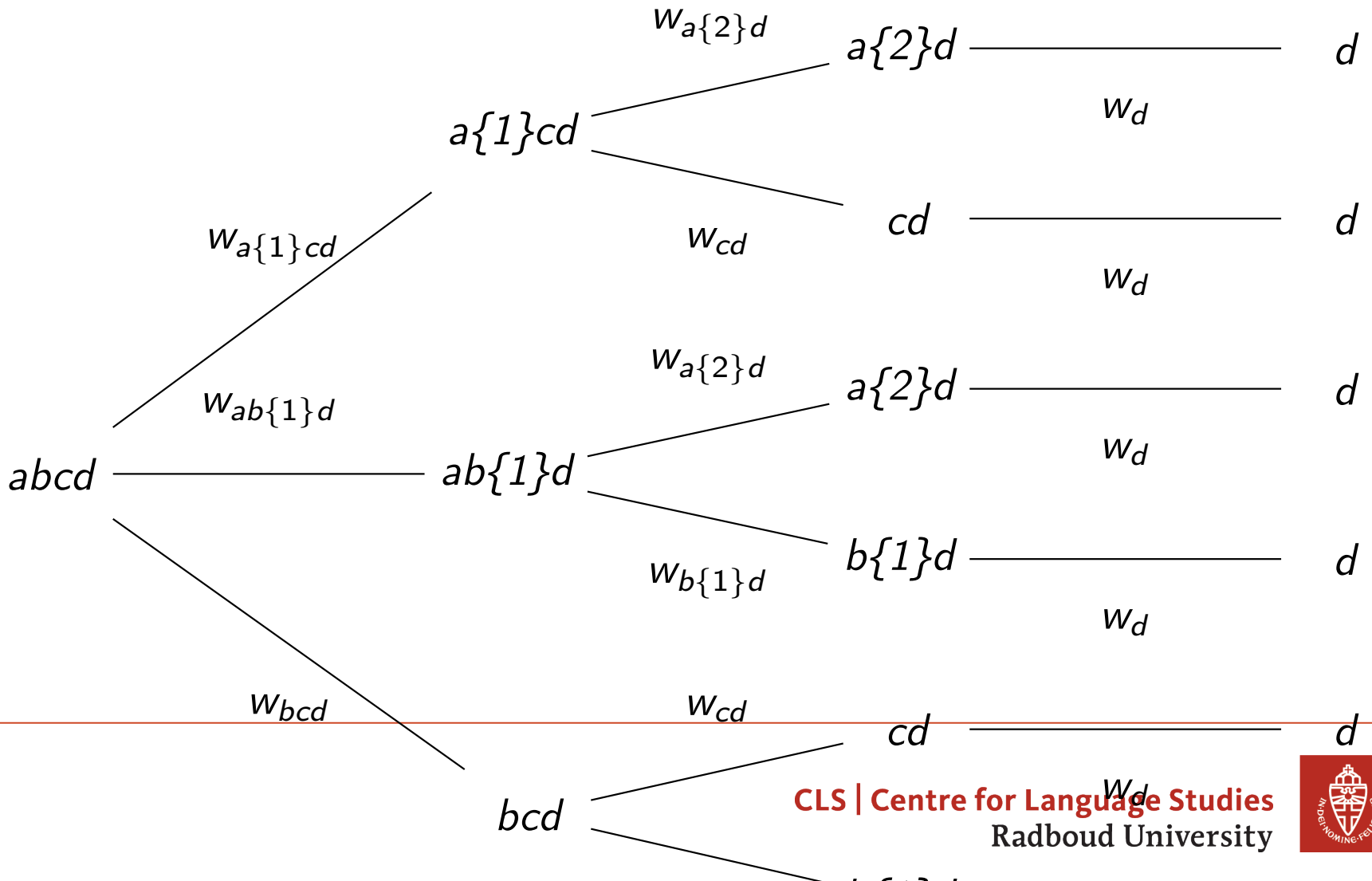
training	1bw				emea			
test	1bw	emea	jrc	wp	1bw	emea	jrc	wp
ngram	129	1 124	941	456	1 761	6	898	1 12
fulluni	125	728	729	392	1 394	6	773	908
Δ%	3	35	23	14	21	−1	14	19
fullnpref	118	700	694	372	1 306	6	705	853
Δ%	6	4	5	5	6	2	9	6
uni	125	728	729	392	1 394	6	773	908
mle	125	000	000	000	1 931	6	1 015	1 22
count	122	893	885	421	1 681	6	889	1 07
ent	132	794	792	434	1 552	6	881	1 03
ppl	157	1 002	1 027	555	2 007	6	1 218	1 32
Δ%	000	000	000	000	000	000	000	000
uni	125	728	729	392	1 394	6	773	908
1bw-value	115	713	694	366	1 212	6	655	655
emea-value	116	692	686	366	1 221	6	651	805
jrc-value	115	694	685	365	1 373	6	709	890
wp-value	115	696	685	317	1 212	6	654	654

hoi

fullnpref	0.0	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.4	0.5	0.6	0.8	1	1.3	1.7	2.1	2.7	3.5	4.5	5.7	7.4	9.4	12.1	15.6	20
1bw																									
emea																									
jrc																									
wp																									

Table: The perplexity values for different fullnpref preference rates with the 1bw model. The 25 steps were sampled in a log space from $[10^{-1.3}, 10^{1.3}]$. The results show that indeed fullnpref-2.0 was a good first guess, with optimal values somewhere between 2.71 and 4.47, depending on the test set.

doei


abcd
$w_{a\{1\}cd}$
$w_{ab\{1\}d}$
w_{bcd}
a{1}cd
ab{1}d
bcd
$w_{a\{2\}d}$
w_{cd}
$w_{a\{2\}d}$
$w_{b\{1\}d}$
w_{cd}
a{2}d
cd
a{2}d
b{1}d
cd
w_d
w_d
w_d
w_d
w_d
d
d
d
d
d

Perplexities on Dutch data

WER on Dutch data