Question 3

Flag question Marked out of 6.00 Not yet answered

a) Table 1 shows bigram counts of seven words in a text corpus. Table 2 shows the unigram counts for the same words. Calculate bigram probabilities for the following word pairs A and B. Show your calculations. (1p)

A: "want to" B: "eat lunch"

b) Solve bigram probabilities for following word pairs A and B that do not have bigram examples in the corpus. Apply either one of these two methods: 1. back off to unigram, 2. add-one smoothing. If you use back-off, use back-off weight *bw*=0.1 . Total number of word types in the vocabulary is 2000 and total number of words in the whole text corpus is 500 000. Show your calculations. (2p)

A: "Chinese want" B: "lunch lunch"

Table 1. Bigram counts. Preceding context word is given in the first column and the current word in the first row.

	I	want	to	eat	Chinese	food	lunch
I	9	1072	0	13	0	0	0
want	4	0	780	0	5	9	5
to	3	0	11	855	3	0	14
eat	0	0	2	0	19	2	57
Chinese	2	0	0	0	0	118	1
food	17	0	15	0	0	0	0
lunch	6	0	0	0	0	1	0

Table 2. Unigram counts.

I	3442
want	1212
to	3123
eat	920
Chinese	199
food	1405
lunch	450

c) What are the weaknesses of the maximum likelihood method you used in a) and b)? (1p)

d) Describe in detail (including updated probability estimate) one method to alleviate these weaknesses. (2p)

Maximum file size: 400 MB, maximum number of files: 1

Files

You can drag and drop files here to add them.

Previous page Next page