T-61.184 Seminar on Statistical (Adaptive) Language Modeling

Exercises for chapter 6: N-grams

Patrik Hoyer, 25.10.2001

1. From a large text corpus you have managed to derive the following bigram model, where Pij = P(word in column j follows word in row i):

	this	exercise	is	stupid
this	0.02	0.80	0.10	0.08
exercise	0.01	0.01	0.97	0.01
is	0.30	0.02	0.01	0.67
stupid	0.31	0.65	0.03	0.01

   [Well, OK, the corpus isn't very large as there are only four words...]

   Given that the first word of a 'test set' is 'this', what is the probability that the 4th word is 'stupid'? What is the probability that the N-th word is 'stupid' in the limit when N approaches infinity? Is this the same as the unigram probability for 'stupid'?

2. The Witten-Bell discounting algorithm can sometimes give higher probability to a word type that hasn't been seen than to a type that has been seen! Consider the following example: Assume that there are D words in our dictionary, and all but one have been seen, and each one only once. That is, we have seen (D-1) types in (D-1) tokens. Using Witten-Bell discounting, what is the probability of the unseen type? What are the probabilities for the other (D-1) types? Compare the two and try to explain the results.