Decoding (CS 344)

Greedy Decoding

Generate one complete translation. At each step, use the single most likely token. Compute the total log probability by taking the sum of the logprobs for each token.

Our translation (logprob = ______): ______ ______ ______ ______ ______ ______

Sampling Decoding

Generate one complete translation. At each step, sample from the top tokens according to their probability. *To do this, pick a random number between 0 and 1, and find the first number under the* ***cumulative probability*** *column that is less than it.*

Then repeat the process again, drawing different random numbers. (If you don't end up with different choices within the first few tokens, re-draw until you do.)

Our translations:

1. (logprob = ______): ______ ______ ______ ______ ______ ______
2. (logprob = ______): ______ ______ ______ ______ ______ ______

Beam Search Decoding

Generate 2 complete translations. Start by taking the top 2 starting tokens. For each of them, find the most likely *following* token. But instead of keeping all 4 possible sequences, only keep the sequences with the largest total logprob (including the new token).