LNG Worksheet: pos tagging

November 20, 2013

Pos-tagging: toy model

You are given the corpus below, where the format word/tag indicates a word with an associated part of speech tag and we restrict ourselves to 4 part-of-speech tags (D=determiner, N=noun,A=adjective,V=verb).

A/D clever/A adult/N will/V train/V children/N

The/D train/N stops/V

An/D adult/A traveller/N knows/V the/D main/A stops/N

On top of the 4 part-of-speech tags mentioned we also include a START for start of sentence and END for end of sentence tag.

Throughout this exercise, do not distinguish between upper-cased and lower-cased versions of the same word.

1. Construct a word-tag table from our corpus as needed for Markov model tagging. Use all words in the corpus and all tags given above.

2. Construct a tag-tag bigram table from the corpus as needed for Markov model tagging. Use all tags (including the start and end tags).

3. We also have a target sentence

   S: The adult stops the train

   What are all the possible tag sequences for this sentence, using our tagset? How many are there? What is the correct tag sequence?

4. From the word-tag table, calculate the probabilities $p(word|tag)$ for all 16 combinations of words in S and the 4 part-of-speech tags. Most of them will be 0, so it should not be so much work.

5. What is the probability a unigram tagger assigns to each possible tag sequence? Which final tag sequence does it assign to the target sentence? Remember that a unigram tagger always assigns the most frequent POS per word in the end.

6. From the tag-tag table, calculate the probabilities $p(tag2|tag1)$ for all bigram tag sequences that occur in any of the possible tag sequences for S. You will need to include the START and END tags.

7. Using the probabilities calculated in the previous questions and the (bigram) Markov model tagging formula, compute the probabilities for all possible tag sequences for S. Which is the most likely?