NLP Worksheet: n-gram modelling and entropy

November 20, 2013

1 n-gram modelling: toy model

You are given the corpus below:

> The cat chases the dog
>
> The dog bites the young cat
>
> The man chases a young dog

1. Using the words *the, dog, bites, a, young, man* as vocabulary write down a bigram model using the above corpus.
2. Using a bigram model, compute the probability of the sentence *The dog bites a man.* If you encounter zero probabilities, use Laplace smoothing to recreate the probabilities.
3. What is the per-word entropy of the above corpus? What is the per-letter entropy?

2 n-gram modelling: realistic model

From the coursework you have a bigram model from the Brown corpus as well as a bigram frequency distribution.
Compute the (unsmoothed) bigram probability of the sentence *The dog bites a man.* from the corpus. (You will need to also use a unigram model for this task; see lab sheet 1).

3 Entropy and the twenty questions game

You have a coin and throw it n times but are not allowed to see the outcome. What is the average number of questions $H_0(X)$ you need to ask to find out the outcome, using best strategy. You are allowed to ask OR questions (*Is the outcome X or Y?*) and you will only get Yes/No answers.

- What is the answer for an unbiased coin?
- What is the answer for a completely biased coin ($p(H) = 0$)?
- What is the answer for a coin with $p(H) = 0.75$?