EMNLP Tutorial 1

Philipp Koehn

This tutorial deals with estimating probabilities and very simple language models. You will need a calculator (with logs).

1. Suppose you are given the following training set:

   the quick brown fox jumps over the lazy dog

   Ignoring spaces, estimate an unconditional distribution over single charachters $p(c)$, where c is drawn from the set of all letters. Let us call this **Model 1**.

2. The log-likelihood of examples $c_1, ..., c_n$ is:

   $$LL = \sum_i \log p(c_i)$$

   Compute the log-likelihood of the training material with respect to **Model 1**. Why is it not 0?

3. What is the log-likelihood of the following testing material

   mary had a little lamb

   How does this compare to log-likelihood for the training material?
   How can you account for sentence length?

4. Create a new **Model 2** that has a higher log-likelihood on the testing material. Does the log-likelihood on the training material improve as well?

5. Compute the perplexity of **Model 1** and **Model 2**, both on the training as well as the test data.

6. Discuss how language models over letters can be used for **language detection**: Given a text, we want to know, which language it is in.

note: based on a tutorial by Miles Osborne