1 Introduction

In this exercise, we are going to start looking at the problem of part-of-speech tagging. More precisely, we are going to build \( n \)-gram models over part-of-speech tags and evaluate them using cross-entropy. This means that we are going to replicate many of the things we did in the language modeling exercise, but for data that have quite different distributional properties. Next time we are going to add a lexical model, so that we can do real part-of-speech tagging.

2 Data and preprocessing

In this exercise, we are going to use the Stockholm-Umeå Corpus (SUC) of Swedish, divided into a training set `suc-train.txt` and a test set `suc-test.txt`. The format is essentially the same as before, with one token per line and blank lines between sentences, but we now have two tab-separated columns, one for the word token and one for its part-of-speech. In this exercise, we are only going to use the part-of-speech tags, which have been extracted in the files `suc-train.tag` and `suc-test.tag`.

**Note on tag set:** We are only using the 25 base tags of the SUC tag set, without any morphosyntactic features.\(^1\)

3 Using MLE for tag language models

Use the program `ngram-mle.py` to estimate \( n \)-gram models from `suc-train.tag`, then use `entropy.py` to evaluate them on `suc-test.tag`. All in all, you should (at least) run the following experiments:

1. Train a unigram model on `suc-train.tag`, test it on `suc-test.tag`
2. Train a bigram model on `suc-train.tag`, test it on `suc-test.tag`
3. Train a trigram model on `suc-train.tag`, test it on `suc-test.tag`

What results do you get? How do they compare to the language modeling results from the previous exercise? Discuss with your group and with your teacher.

4 Add smoothing

Use your implementation of additive smoothing from the previous exercise to train and evaluate smoothed versions of the unigram, bigram and trigram models. Compare the results both to the MLE results and to the smoothed results for language modeling. What similarities and differences do you find? Can you explain them?

\(^{1}\)For more information on the SUC tag set, see [http://spraakbanken.gu.se/parole/Docs/sucnyckel.suc2](http://spraakbanken.gu.se/parole/Docs/sucnyckel.suc2).