Computational Linguistics

Exercise on Conditional Random Fields

In this exercise you should build a POS-tagger.

In the first part (due July 4th) you should make yourself familiar with the task and the software.

1. Please download the CRF++ software from http://crfpp.sourceforge.net/ .
2. Run ./configure and afterwards make to compile the software.
3. Download the data (http://www.lsv.uni-saarland.de/download/pos_data.tar)
4. Train a tagger that uses the present word as a feature. Use 1%, 2%, 5& 10%... of the available data to train different models.
5. Measure number of wrongly assigned tags on the development corpus
6. Plot the amount of training data used vs. the number or errors

Please present your results on July 4th using one or two slides.

In the second part, you should try to come up with additional features and explore all the possibilities of CRF++. Try to build the best possible tagger you can.

Before the exercise session on July 7th, please send me a link to the input file for your training containing all the features as additional columns, the template file, your trained model (output of CRF++) and a script that runs the feature generation on unseen test data. Your answer should also contain a short report.

In the exercise session on July 7th, please present our findings using a few slides.