Computational Linguistics

Exercise on Conditional Random Fields

In this exercise you should build a chunker.

In the first part (due 31 May) 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 training (http://www.lsv.uni-saarland.de/download/wsj_train.chunk.corp) and development (http://www.lsv.uni-saarland.de/download/wsj_dev.chunk.corp)
4 Train a chunker 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 chunk tags on the development corpus
6 Plot the amount of training data used vs. the number or errors

Please present your results on May 31 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 chunker you can.

Before the exercise session on June 10th, please send me a link to the input file for your training containing all the features as additional columns, the template file and your trained model (output of CRF++). Your answer should also contain a short report.

In the exercise session on June 10th, please present our findings using a few slides.

Note:

The best chunker as evaluated on unseen test data will get a *Buchgutschein* (book voucher) for 20 €. I will bring the test data and the voucher to the last exercise session.