

Computational Linguistics

Exercise on Conditional Random Fields

Due: 18.6.

In this exercise you should build a part-of-speech tagger.

- 1 Please download the CRF++ software from <http://crfpp.sourceforge.net/> .
- 2 Run `./configure` and afterwards `make` to compile the software.
- 3 Download training (train.corp) and development (dev.corp) data from the course home page
- 4 Train a tagger that uses the present word as a feature on train.corp. Use 1%, 2%, 5% & 10%... of the available data to train different models
- 5 Measure number of wrongly assigned POS tags on dev.corp.
- 6 Plot the amount of training data used vs. the number of errors

Please send me a link to the input file for your training containing all the features, your template file and your model. Your answer should also contain a description of what you did.

Optional:

Try to design new and better features to get an improved tagger. The best POS tagger will get a *Buchgutschein* (book voucher) for 20 euros.