Exercise: Machine Learning

Research question: What is the relative importance in POS-tagging of

- Knowledge about the focus word
- Knowledge about the left context
- Knowledge about the right context
- Knowledge about non-adjacent words
- The choice of machine learning method

Method: Data

Given that we do not have all that much time, we will focus on words which can be either noun (singular) or verb (infinitive or present tense), such as *labour*. I took all instances of such words from the LOB corpus. For each word, I built a case with

- Features for the focus word (F), the preceding two words (P1, P2), and the next two words (N1, N2). When this window extends beyond the utterance boundary, the empty positions are marked as *<O>.*
- For each word, we take the word itself (W), the potential tags for that word (A; in order of observed frequency and at most three), and the appropriate tag in the current context (D).

The resulting case base can be found in the file lobtlb_nnvb.csv (attached).

Method: Machine learning system

We will use the system Weka for our investigation. Weka gives access to many different ML-techniques. Often it suffices for your needs. However, it also has disadvantages, such as trouble with processing very large case bases.

Weka can be downloaded from http://www.cs.waikato.ac.nz/ml/weka/index.html. Please download and install on your own machine.

Method: Investigation

1. Start Weka, choose Explorer
2. Under the Preprocessing tab, open the file lobtlb_nnvb.csv
3. Use the feature selection on Preprocessing tab to select the features you want for each specific test
    - For the first try, only keep P1D (previous tag) and FD (target class)
    - For later tests, you can go back here and adjust your choice. Removing further features is always possible, but (as far as I can see) if you want to have more features, you have to reload the file.

4. Go to the Classify tab and select a machine learning method.
    - Use at least NaiveBayes (under Bayes) and J48 (under Trees).
    - If you want, you can try other methods. But don't hesitate to use the Stop button if learning takes too long.
5. Use 10-fold cross-validation to see how well the system succeeds at its task with that method and those features.
    - Focus mainly on the percentage of correctly classified cases. In real research you would look at more detail, including an error analysis of where the system fails; now there is no time for that. Exception: when using all features, have a look at how the various features are used in the classification model.
    - Keep notes so you can report the results later
6. Repeat with the various feature and system combinations that you need to answer the research question.
    - Every time return to step 2, 3 or 4 to do the desired classification
    - For this exercise, restrict yourself to P2D, P1D, N1D, N2D, FW, FD (i.e. leave out lexical information about the context). Note that in practice, using left-to-right tagging, you would have N1A and N2A available rather than N1D and N2D. Also note that, normally, FA is extremely valuable, but in this case base it is always NN|VB or VB|NN, and the order can be derived from all cases with the same FW, making it redundant.
    - Taking all possible subsets is exaggerated, e.g. looking at P2D but not P1D is probably not very useful.
7. Organize and report your measurements, and draw conclusions about the research question.
    - You have to make a report, but you do NOT have to put effort into making this a well-written paper.
8. In class, we will compare experiences and conclusions.