EDAN20
Harald Nordgren
900404-1936

Lab Report 3

We downloaded training and test sets.

We glanced at the *CoNLL 2000* shared task baseline, which seemed pretty good. We modified *BLChunkerIncomplete* to collect chunk tags, and to assign part-of-speech values to tags. We then measured performance with *eval.sh*, which had an accuracy total of 77.07%.

We then tried replicating the experiment with Weka, starting with removing the word attribute from *chunk_header.arff* and running *MLChunker*. We loaded the *ARFF* file into Weka, created a classifier and saved the model.

We used *MLChunker* and *WekaGlue* to predict chunk tags, added *weka.jar* to the library and ran the chunker. The results were the same as for the previous Java code.

Finally, we modified *MLChunker* by including a feature vector of size 2. We looped for each word of a sentence, holding on the last word of every round and saving these combinations. The first word of every sentence was tagged with *BOS*. Compiling and rebuilding our decision tree in Weka gave us an accuracy of 81.88% when we ran the program.

It is apparent that using a larger word window gives a better accuracy, but it adds a lot of complexity to the program, and we struggled to implement it correctly.