LING572 Hw9: TBL
Due: 11:45pm on March 11, 2010

The example files are under dropbox/09-10/572/hw9/examples/.

Q1 (40 points): Write a TBL trainer, **TBL_train.sh**, for the text classification task.

- The command line is: TBL_train.sh train_data model_file min_gain
- The initial annotator simply tags each document with the **first** class in the training data (e.g., if the training data is **train2.txt**, the first class would be **"guns"**).
- train_data has the same format as before (i.e., Mallet text format)
- model_file has the default classname (i.e., the first class in the training data) in the first line, followed by a list of transformations (one transformation per line). The transformation line has the format "featName from_class to_class net_gain".
- If the net gain of the best transformation for the current iteration is less than min_gain, the TBL training will stop.
- In order to find the best transformation, I mentioned in class that you needed to go over all the training instances that the current class labels are incorrect. In fact, you need to go over all the instances **including** the ones whose the current class labels are correct.

  If your algorithm is efficent, for every iteration of training, you need to go over the training data only once. The trick is that for each training instance, determine what transformations would be triggered by the instance and update their net gains accordingly.

Q2 (30 points): Write a TBL decoder, **TBL_classify.sh**, that uses a TBL model to classify test instances.

- The command line is: TBL_classify.sh test_data model_file sys_output N
- test_data has the same format as before (i.e., Mallet text format)
- model_file is the model created by TBL_train.sh
- sys_output has the format "instanceName trueLabel sysLabel transformation1 transformation2 ...": trueLabel is the label in the gold standard, sysLabel is the label produced by the TBL classifier, each transformation has the format "featName from_class to_class".
- N is the number of transformations in the model_file that will be used. For instance, suppose the model file has 1000 transformations and N is 10, then only the first 10 transformations in the model file will be used, and the rest will be totally ignored as if they were not in the file.

Q3 (30 points): Run the TBL trainer and classifier with **train2.txt** as the training data and **test2.txt** as the test data.

(a) Fill out Table 1.

(b) What conclusions can you draw from the experiments?

Table 1: The classification results

N	Training Accuracy	Test accuracy
1		
5		
10		
20		
50		
100		
150		
200		
250		

Submission: Submit a tar file via CollectIt. The tar file should include the following.

- The source code for Q1 and Q2.
- The answer to Q3
- The model file created in Q3, and the sys_output file for the test data when all the transformations are used.