LING572 Hw7: TBL
Due: 11pm on Feb 22, 2017

The example files are under dropbox/16-17/572/hw7/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".
- min_gain should be a positive integer. If it is not, the code should print out an error message and exit.
  - If the net gain of the best transformation for the current iteration is less than min_gain, the TBL training will stop.
  - If min_gain is 1, the trainer will not stop until the best transformation in the current iteration cannnot provide a positive gain. In this case, we say that the model file contains all the transformations with positive gains.
- In order to find the best transformation, you need to go over all the instances **including** the ones whose the current class labels are correct.

  If your implementation is efficent, for every iteration of training, you need to go over the training data only once to find the best transformation. The trick is that for each training instance, determine what transformations would be triggered by the instance and update their net gains accordingly. See the slides for hw7.

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 for decoding, 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. N is the number of transformations used by TBL_classify.sh. Please call the model file produced by the trainer with min_gain=1 *model_file*, and the system output file of the classifier *sys_output_N*.

(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 hw7 note file that includes the answer to Q3.
- The source code for Q1 and Q2.
- For Q3, please submit *model_file* and *sys_output_250*. For other N values, submitting *sys_output_N* is optional.