COMS 4705 Project 3 Report
Professor Michael Collins
Qingxiang Jia
qj2125

Please run the following command to grade the project:
```
chmod -x grade.sh
./grade.sh
```

Question 4
My implementation (`EM1.java`) has the output that is identical to the sample output. To see the output file, please open file `q4_output.txt`. For more details on implementation, please see the source code in-line comments.

Question 5
Notice that my implementation (`EM2.java`) does not run IBM model 2 again, instead, it reads in the serialized file `t.ser` that is produced in Question 4. The output file has been named as `q5_output.txt`.

The two models result in similar but slightly different alignment. Because we have modified the parameter estimation process. In particular, we added the alignment information conditioned on the sentence length; consequently, the calculation of δ has changed too. The alignment should be more accurate than the one generated in Question 4.

Question 6
For this problem, I directly use the deserialized file `em2_t.ser` and `em2_q.ser` to avoid unnecessary computation. The efficiency of this part (`Aligner.java`) is not so great since for each word of each German sentence, we need to compute the value for every English sentence. For each German token, we need to try all possible alignments to find the maximum. However, since we do not have a large data set, and I have applied some optimization (converting everything into primitive-like `int[]` in java speeds up the whole process significantly), the process takes around 30s only. The only problem I had during unscrambling is that due to my implementation of IBM model 2, I need to make sure that certain sentence length pair (l, m) in the scrambled.en is in my parameter t and q, since the sentences in `scrambled.en` are unseen. This is also true for unseen words checking.