Homework 1: Solutions and Notes on Grading

Junkun's Notes Revised from a Previous Rubric
CS 539

October 8, 2019

Total points possible: 12.

1. Possible points: 0.75

   The dictionary contains 105,403 words, 27 character types, and 773,991 character tokens. In Fig 1, we ignore the labels of states.

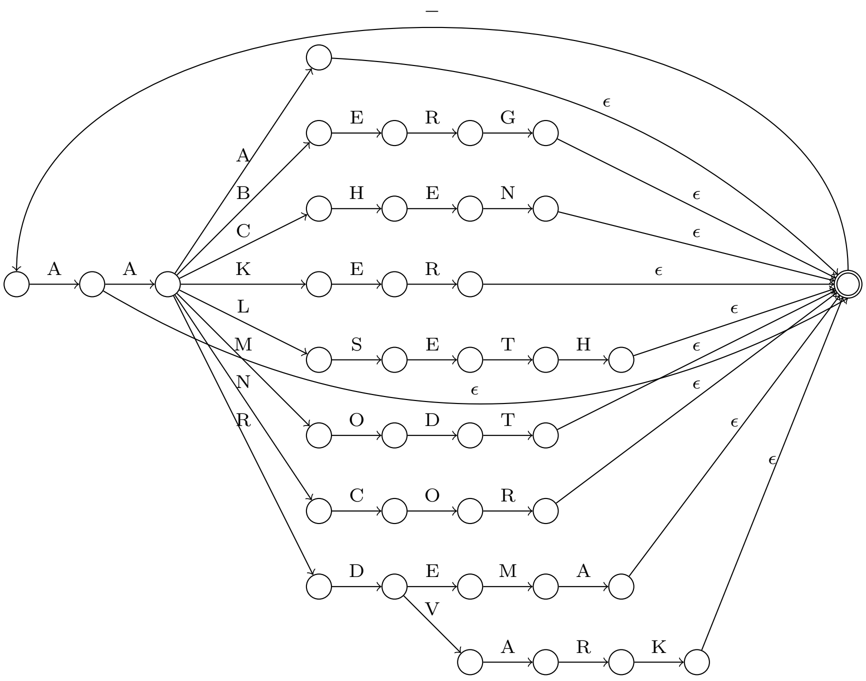


Figure 1: small fsa

2. Possible points: 3.0. The basic solution was to use a prefix trie. Depending on how you handled the looping back to the root, this gives 256,332 or 175,778 states.

   1 baseline point for the script/fsa file.

   a. 1 point based on the number of states.

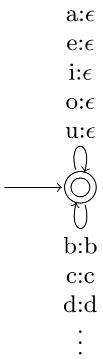
   b. 1 point for showing the output of Carmel on the first five lines of each file.

For part a. we assigned points as follows:

states	points
Roughly 250,000 or less	1
A lot more than 250,000	less than 1

3. Possible points: 2.25.

   a. 0.75 points



   b. 0.75 points for showing the output, and saving the output to `strings.novowels`

   c. 0.75 points for listing the naively recovered output.

4. Possible points: 1.5. Typical accuracy was 1.3%.

   a. 0.5

   b. 0.5

   c. 0.5

5. Possible points: 2.25. Use composition to combine the FSA from Question 2 and Question 3. Typical accuracies were in the low 30s. Many people did the composition themselves instead of using Carmel to do it automatically. There was no penalty or bonus for this, but using Carmel is certainly much easier.

   a. 0.75

   b. 0.75

   c. 0.75

6. Possible points: 2.25. The typical solution was to collect word frequencies from `strings`, which gives an accuracy of about 91.3.

   a. 0.75

   b. 0.75

   c. 0.75

   For part c. we assigned points as follows:

accuracy	points
90–92	0.75
85–90	0.60
60–85	0.45
45–60	0.30
30–45	0.15

It was fairly easy to get good results overfitting a character ngram model from the `strings.txt` file. So for anyone who also attempted a word-to-word bigram model, trained it on another large corpus (without overfitting), and reported this in their pdf got 1 point of extra credit, as long as their bigram performance would be lower than 0.90 accuracy.