Natural Language Processing Homework 7

Katie Chang + Willis Wang

April 29, 2016

README

1   Q2

1.1   first

First accepts a language that starts with at least one 0 and ends with at least three 1s. There can be any combination of 0s and 1s in between the leading 0 and the trailing 111. Since the ? denotes optionality, the fourth 1 is not necessary for the string to be accepted in this language.

ii. This language is a subset of the language Bit*.

The 0 and 1 are quoted in the .grm file because these two FSMs are defining specific strings that this language will use. Essentially, the quotes indicate the initializing of the letters of the alphabet that this language will use.

iii. There are 5 states and 9 arcs in this FST.

1.2   b. second

i. export Second = Optimize[Zero Bit* One One One];

ii. If First and Second are equivalent, then Disagreements should accept nothing.

Running fstinfo on Disagreements.fst, we can see that this FST has 0 states and 0 arcs, and in fact has a 0 count of everything. The initial state value is -1. Therefore, it should be unable to accept any sort of string.

1.3   c

i. First.fst now has 20 states and 25 arcs. Second.fst now has 13 states and 17 arcs. Disagreements.fst has 68 states and 88 arcs (!!!).

  ii. The two separate sub-FSAs is because the Disagreement FST is taking the union of two FSTs, First - Second and Second - First.

  iii. The results are the same. The only difference there is is the size of the FST and topology.

1.4   d

After creating Disagreements.fst, the number of states and arcs is the same as before. This shows that it doesn't matter whether or not we optimize the sub-FSTs, but there would be a lot of redundant work done. We should optimize if we want to avoid excessive overhead costs.

2   Q3

See binary.grm.

3   Q4

3.1   a

Input language: "a"("b"* — "c"+ — "")"a"

3.2   b

"abca" has no outputs.

  "aba" has 1 output = "axa"

  "aa" has 2 outputs = "aa" and "africa"

  "aca" has more than 2 outputs.

3.3   c

Takes all strings that start and end with one a, as well as one of the following criteria:

1. replaces all "b"s between the two "a"s with "x"s
2. replaces all "c"s between the "a"s with any number of "y"s
3. replaces any "" between the "a"s with either the empty string or "fric".

3.4   d

Consistent. 10 states and 16 arcs.

4   5

4.1   c. Parity1

This particular transducer does not take into account the empty string.

4.2   f. UnParity

UnParity inverts the transduction. The only acceptable strings are "0" and "1" / any number of 0s before one "1". The former refers to even binary numbers, and the output lists a bunch of random ones. The latter refers to odd binary numbers, and will list a bunch of them.

5   6

Used overleaf.com to generate LaTeX document.