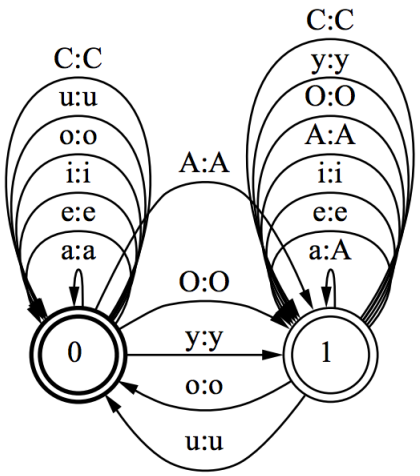
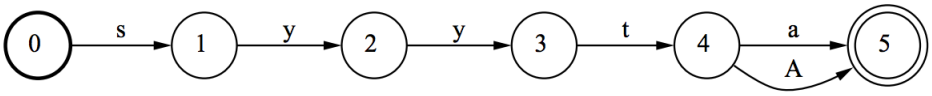
CMPT 413/825 - Spring 2014 - Midterm

There are three questions (20pts total). Please write down "Midterm" on the top of the answer booklet.
When you have finished, return your answer booklet along with this question booklet.

(1) (10pts) **Finite-state transducers:** The following finite-state transducer (FST) *V* implements basic Finnish vowel harmony. The symbol A stands for ä, O stands for ö and C stands for a consonant letter, which is any letter that is not in the set of vowels, { a, e, i, o, u, y }.



a. (5pts) The following FSM *f* represents two forms of the Finnish word *syy+tä* (*reason*). Again, A stands for ä.



Id(*f*) is an FST that recognizes a pair of strings (x, x) where x is a string recognized by FSM f. Provide the result of first composing *Id*(*f*) with the FST *V* shown above and then projecting the output side of the composed FST, i.e. provide *project_output*(*Id*(*f*) ∘ *V*).

Answer: *project_output*(*Id*(*f*) ∘ *V*) is the FSM that recognizes the string *syttA* or syttä.

b. (5pts) Provide a rewrite rule that is equivalent to the FST *V* shown above. Assume it is left to right, iterative and obligatory.

Answer:

```
a -> A / [A,O,y] C* ([i,e] C*)* ________
```

(2) (5pts) **Edit distance**: Assume insertion of a character has cost 1, deletion has cost 1, and substitution of one character for another has cost 2.

a. (2pts) What is the minimum edit distance value between target word *goal* and source word *hole*?

b. (3pts) The following is a visual display of one possible alignment between target word *goal* and source word *hole* using the usual notation.

```
g o a l _
  |   |
h o _ l e
```

Using the above notation for alignments, provide any other possible alignments that have the same edit distance as the alignment shown above.

Answer:

```
levenshtein distance = 4
alignment number 1 for [4,4]:
_ g o a l _
    |   |
h _ o _ l e

alignment number 2 for [4,4]:
g _ o a l _
    |   |
_ h o _ l e

alignment number 3 for [4,4]:
g o a l _
  |   |
h o _ l e

total of 3 alignments
```

(3) (5pts) **Language Modeling**

Consider a language model over character sequences that computes the probability of a word based on the characters in that word, so if word $w = c_0, c_1, \ldots, c_n$ then $P(w) = P(c_0, \ldots, c_n)$. Let us assume that the language model is defined as a bigram character model $P(c_i \mid c_{i-1})$ where

$$P(c_0, \ldots, c_n) = \prod_{i=1,2,\ldots,n} P(c_i \mid c_{i-1})$$

Katz backoff smoothing is defined as follows:

$$P_{katz}(c_i \mid c_{i-1}) = \begin{cases} \frac{r^*(c_{i-1},c_i)}{r(c_{i-1})} & \text{if } r(c_{i-1}, c_i) > 0 \\ \alpha(c_{i-1}) P_{katz}(c_i) & \text{otherwise} \end{cases}$$

where $r(\cdot)$ provides the (unsmoothed) frequency from training data and $r^*(\cdot)$ is the Good-Turing estimate of the frequency r.

Provide the equation for $\alpha(c_{i-1})$ that ensures that $P_{katz}(c_i \mid c_{i-1})$ is a proper probability.

Answer:

Step by step derivation below. We are just looking for the end result.

$$\begin{aligned}
\sum_{c_i}\left(\frac{r^*(c_{i-1},c_i)}{r(c_{i-1})}+\alpha(c_{i-1})P_{katz}(c_i)\right) &= 1\\
\sum_{c_i:r(c_{i-1},c_i)>0}\frac{r^*(c_{i-1},c_i)}{r(c_{i-1})}+\alpha(c_{i-1})\sum_{c_i:r(c_{i-1},c_i)=0}P_{katz}(c_i) &= 1\\
\alpha(c_{i-1})\sum_{c_i:r(c_{i-1},c_i)=0}P_{katz}(c_i) &= 1-\left(\sum_{c_i:r(c_{i-1},c_i)>0}\frac{r^*(c_{i-1},c_i)}{r(c_{i-1})}\right)\\
\alpha(c_{i-1}) &= \frac{1-\left(\sum_{c_i:r(c_{i-1},c_i)>0}\frac{r^*(c_{i-1},c_i)}{r(c_{i-1})}\right)}{\sum_{c_i:r(c_{i-1},c_i)=0}P_{katz}(c_i)}\\
\alpha(c_{i-1}) &= \frac{1-\left(\sum_{c_i:r(c_{i-1},c_i)>0}\frac{r^*(c_{i-1},c_i)}{r(c_{i-1})}\right)}{1-\left(\sum_{c_i:r(c_{i-1},c_i)>0}P_{katz}(c_i)\right)}
\end{aligned}$$

Also acceptable is the somewhat less precise answer which assumes $\sum_{c_i} P_{katz}(c_i) = 1$:

$$\alpha(c_{i-1}) = 1-\sum_{c_i}\frac{r^*(c_{i-1},c_i)}{r(c_{i-1})}$$