CMPT 413 - Spring 2011 - Midterm #1

Please write down “Midterm #1” on the top of the answer booklet.

When you have finished, return your answer booklet along with this question booklet.

(1) (12pts) **Minimum Edit Distance**

You are given a visual display of minimum distance alignments for three examples. From these examples, derive the costs for the operations of insertion, deletion and replacement in the target word. Assume that matching between characters that are unchanged is a zero cost operation.

The 1st line of the visual display shows the *target* word and the 3rd line shows the *source* word. An insertion in the target word is represented as an underscore in the 3rd line aligned with the inserted letter in the 1st line. Deletion from the source word is represented as an underscore ‘_’ in the 1st line aligned with the corresponding deleted character in the source on the 3rd line. Finally, if a letter is unchanged between target and source then a vertical bar (the pipe symbol ‘|’) is printed aligned with the letter in the 2nd line.

```
min edit distance = 1.25
g a r g l i n g
| |     |
g a m b l _ _ e

min edit distance = 0.75
g a m b l e
|       | |
g o o g l e

min edit distance = 1.45
u n g a r g l _ _ e
    |     | |
_ _ g o o g l i n g
```

Provide the cost of letter insertion, the cost of letter deletion and the cost of letter replacement.

Answer:

```
replacecost = 0.25;
equalcost   = 0;
insertcost  = 0.25;
deletecost  = 0.1;
```

(2) (24pts) **Rewrite Rules**

The rewrite rule $v \rightarrow cvc/cv^{*}\_c$ is applied on an input string from left to right, iteratively. The application of the rewrite rule should be constrained in the usual way so that it is equivalent to a regular relation.

a. Provide the step by step application of the rewrite rule on the input string *cvc*. Show the lhs and the left and right context match for each step.

b. Provide the step by step application of the rule on the input string cv^3c (which is a concise way to write the string *cvvvc*). Show the lhs and the left and right context match for each step.

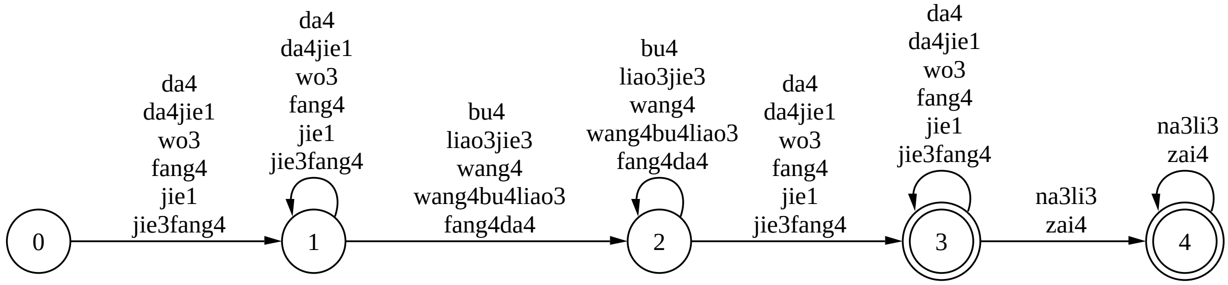
c. For the input language $\mathcal{I} = \{cv^nc \mid n \geq 1\}$, provide the output language produced by the application of the rewrite rule.

Answer:

The lhs of the rewrite rule where left and right context matches is written as v:

1. $c\mathsf{v}c \rightarrow ccvcc$
2. $cv^3c = cvv\mathsf{v}c \rightarrow cvvc\underline{v}cc$. Note that at this point $\underline{\text{v}}$ cannot be rewritten even though the left and right context matches, because of the restriction that prohibits rewriting any string that was inserted. This restrictions permits the rule to generate only regular relations. Then $cv\mathsf{v}c\underline{v}cc \rightarrow c\mathsf{v}c\underline{v}cc\underline{v}cc \rightarrow cc\underline{v}cc\underline{v}cc\underline{v}cc = c(cvc)^3c$.
3. Output language after application of the rewrite rule is $\{c(cvc)^nc \mid n \geq 1\}$

(3) (24pts) In this question, we will use finite state transducers (FSTs) for word segmentation in Chinese. The finite state machine below is an extremely simple grammar for sentences in Chinese using this lexicon. However, this FSM does not generate spaces between the words. For example, this FSM generates the string: *wo3wang4bu4liao3jie3fang4jie1zai4na3li3*.



The lexicon used by this FSM is given in the table below.

Chinese word (pinyin)	English translation
da4	big
da4jie1	avenue
wo3	I
fang4	place
jie1	avenue
jie3fang4	liberation
bu4	not
liao3jie3	understand
wang4	forget
wang4bu4liao3	unable to forget
fang4da4	enlarge
na3li3	where
zai4	at

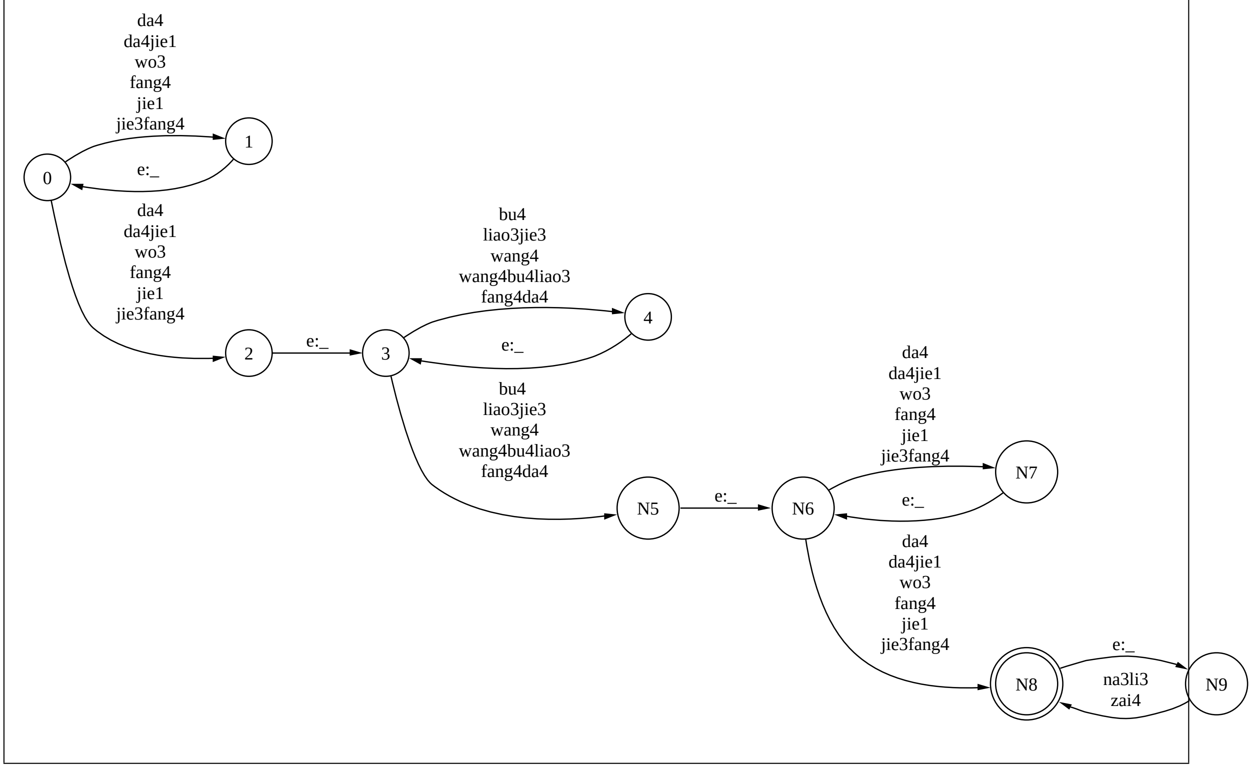
Table 1: Lexicon for Question 3.

a. Construct an FST T_{3a} where the input is a Chinese sentence with no spaces between words as accepted by the FSM above and the output has a space between any two words defined in Table 1, thus segmenting the input sentence into words. Denote the space character as ␣ in your FST.

You can use the following symbols to save time when drawing your FST:

Symbol	Value
A	`da4 da4jie1 wo3 fang4 jie1 jie3fang4`
B	`bu4 liao3jie3 wang4 wang4bu4liao3 fang4da4`
C	`na3li3 zai4`

Answer:



b. Use the following Chinese sentence without spaces between words as an input to T_{3a} and produce at least two different word segmentations (that is: two sentences with spaces between words corresponding to different word segmentations):
wo3wang4bu4liao3jie3fang4jie1zai4na3li3

Answer:

- *wo3 wang4 bu4 liao3jie3 fang4 jie1 zai4 na3li3*
- *wo3 wang4bu4liao3 jie3fang4 jie1 zai4 na3li3*

c. Using the English entries in Table 1, provide one word segmentation of the Chinese sentence from Question 3b that is closest in meaning to the following English sentence:
I was unable to forget where Liberation Avenue is.

Answer:

- *wo3(I) wang4bu4liao3(unable to forget) jie3fang4(Liberation) jie1(Avenue) zai4(at) na3li3(where)*