CS221 Fall 2018 - 2019 Homework 3

Name: Dat Nguyen
Date: 2/12/2019

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1: word segmentation

(a) Query string: "Therearecountlessanimalsonearth".
Correct segmentation: "There are countless animals on earth".
Wrong segmentation: "There are count less animals on earth".
Where 1-gram and c("count") = 0.1, c("less") = 0.2 and c("countless") = 0.2

Problem 2: vowel insertion

(a) Query string: "th bs r n btfl flwr"
Correct vowel insertion: "the bees are on a beautiful flower".
Wrong vowel insertion: "the books are on a beautiful flower".
Where bigramCost('the', 'bees') = 0.2, bigramCost('bees', 'are') = 0.1, bigramCost('the', 'books') = 0.1, and bigramCost('books', 'are') = 0.3.

Problem 3: putting it together

(a)
- States: list of actions made in each step, where each action includes 2 kinds of information: next index to insert space after that in the query string and the choosen word based on the substring between most recent 2 spaces of the query string.
- Actions: as explained above.
- Cost: the bigram cost of the previous choosen word and the currently choosen word.
- Initial state: empty list.
- End state: if the next index to insert space after that is the last index of the query string.

(c) We define $u_b(w) = \underset{w' \in Vocabulary}{min} b(w', w)$, and the following relaxed problem.

- States: like the original problem.

- Actions: like the original problem.

- Cost: let sb be substring between most recent 2 spaces of the query string resulting from state s and action a. We define the cost to be $Cost_{rel}(s,a) = \min_{st \in possibleFills(sb)} u_b(st)$.

- Start state: like original problem.

- End state: like original problem.

Next we will prove that $Cost_{rel}(s,a) \leq Cost(s,a)$. Let sb be substring between most recent 2 spaces of the query string and st' be the choosen word from possibleFills(sb) resulting from state s and action a. We have

$$\begin{aligned}
Cost_{rel}(s,a) &= \min_{st \in possibleFills(sb)} u_b(st) \\
&\leq u_b(st') \\
&= \underset{w' \in Vocabulary}{min} b(w', st') \\
&\leq b(\text{prevWord}, st') \\
&= Cost(s,a)
\end{aligned}$$

Therefore if we let $h(s) = FutureCost_{rel}(s)$ then $h(s)$ is consistent. Since the cost for the relaxed problem are all equal for choosen words given substring between most recent 2 spaces of the query string, we can cosider the following problem to save computation.

- States: list of next index to insert space after that in the query string

- Actions: next index to insert space after that in the query string

- Cost: let sb be substring between most recent 2 spaces of the query string resulting from state s and action a. We define the cost to be $Cost_{rel}(s,a) = \min_{st \in possibleFills(sb)} u_b(st)$.

- Start state: empty list.

- End state: if the next index to insert space after that is the last index of the query string.

And when doing in the original problem we let $FutureCost(s, a)$ be the future cost in the relaxed problem starting from the current substring in the query string being considered by the original problem.

(d) UCS is a special case of A^* if we let $h(s) = 0$.
BFS is a special case of UCS if we let every edge have cost $= 1$.