Edit Distances: Verification

Yufei Tao
KAIST

June 13, 2013
Given two strings $s$, $t$, we already know how to compute their edit distance $edit(s, t)$ using dynamic programming in $O(|s||t|)$ time. It turns out that we can do better if we only need to verify whether $edit(s, t) \leq d$. This can be done in $O(|s| + |t| + d \cdot \min\{|s|, |t|\})$ time.

For simplicity, we will assume $|s| = |t| = \ell$. It is left as an exercise for you to extend our discussion to the case of $|s| \neq |t|$.

Our goal now is to verify whether $edit(s, t) \leq d$ in $O(d\ell)$ time for $d < \ell$ (if $d \geq \ell$, the answer is trivially yes).
Recall that, in order to compute $edit(s, t)$ in $O(\ell^2)$ time, our strategy was to fill in an $(\ell + 1) \times (\ell + 1)$ array $A$. To solve the verification problem, we will adopt a similar strategy, except that we will fill in only a hexagon part of $A$, as explained next.
Let us first define the **gray boundary cells** to be

- At row 0, the left most $d$ cells.
- At column 0, the top most $d$ cells.

Define the **blue boundary cells** to be

- At row $\ell + 1$, the right most $d$ cells.
- At column $\ell + 1$, the bottom most $d$ cells.

An example with $\ell = 8$ and $d = 2$:
Define the **yellow boundary cells** to be:

- $A[0, d], A[1, d + 1], ..., A[\ell + 1 - d, \ell + 1]$
- $A[d, 0], A[d + 1, 1], ..., A[\ell + 1, \ell + 1 - d]$

An example with $\ell = 8$ and $d = 2$: 

![Diagram](image)
Define the green cells to be all those cells inside the region surrounded by the gray yellow, and blue boundary cells.

An example with $\ell = 8$ and $d = 2$: 

![Diagram illustrating the green cells within the specified region.](image-url)
We fill in only the colored cells (i.e., ignoring the others) as follows:

1. Fill in the gray cells normally.
2. Put $d + 1$ in all the yellow cells.
3. Compute the green and blue cells in the same manner as in the $O(\ell^2)$-time algorithm (i.e., row by row, and left to right at each row).

Report yes if $A[\ell + 1, \ell + 1] \leq d$, and no, otherwise.

Since there are only $O(d\ell)$ colored cells, the running time is $O(d\ell)$. 
Example: $s = \text{humanity}$, $t = \text{hunamity}$, and $d = 2$.

After the first two steps:
Example: $s = \text{humanity}, t = \text{hunamity}, \text{and } d = 2$.

After all steps:

So we conclude $\text{edit}(s, t) \leq 2$. 
Think

Why is the algorithm correct?