Example

- Wind has blown away the +, *, (, ) signs
- What’s the maximal value?
- Minimal?

```
2 1 7 1 4 3
```

```
• 2 1 7 1 4 3
• (2+1)*7*(1+4)*3 = 21*15 = 315
• 2*1 + 7 + 1*4 + 3 = 17
```

```
• Q: How to maximize the value of any expression?

2 4 5 1 9 8 12 1 9 8 7 2 4 4 1 1 2 3 = ?
```

Dynamic programming

- Avoid calculating repeating subproblems
- \( \text{fib}(1) = \text{fib}(0) = 1; \)
- \( \text{fib}(n) = \text{fib}(n-1) + \text{fib}(n-2) \)

- Although natural to encode (and a useful task for novice programmers to learn about recursion) recursively, this is inefficient.
Structure within the problem

- The fact that it is not a tree indicates overlapping subproblems.

- A dynamic-programming algorithm solves every subsubproblem just once and then saves its answer in a table, thereby avoiding the work of recomputing the answer every time the subsubproblem is encountered.

Topp-down (recursive, memoized)

- **Top-down approach**: This is the direct fall-out of the recursive formulation of any problem. If the solution to any problem can be formulated recursively using the solution to its subproblems, and if its subproblems are overlapping, then one can easily memoize or store the solutions to the subproblems in a table. Whenever we attempt to solve a new subproblem, we first check the table to see if it is already solved. If a solution has been recorded, we can use it directly, otherwise we solve the subproblem and add its solution to the table.

Bottom-up

- **Bottom-up approach**: This is the more interesting case. Once we formulate the solution to a problem recursively as in terms of its subproblems, we can try reformulating the problem in a bottom-up fashion: try solving the subproblems first and use their solutions to build-on and arrive at solutions to bigger subproblems. This is also usually done in a tabular form by iteratively generating solutions to bigger and bigger subproblems by using the solutions to small subproblems. For example, if we already know the values of $F_{41}$ and $F_{46}$, we can directly calculate the value of $F_{42}$.

Dynamic programming is typically applied to **optimization problems**. In such problems there can be many possible solutions. Each solution has a value, and we wish to find a solution with the optimal (minimum or maximum) value.

- We call such a solution an optimal solution to the problem, as opposed to the optimal solution, since there may be several solutions that achieve the optimal value.

The development of a dynamic-programming algorithm can be broken into a sequence of four steps.

1. Characterize the structure of an optimal solution.
2. Recursively define the value of an optimal solution.
3. Compute the value of an optimal solution in a bottom-up fashion.
4. Construct an optimal solution from computed information.
Matrix multiplication

- \( \text{for } i = 1..n \)
- \( \text{for } j = 1..k \)
  
  \[ c_{ij} = \sum_{x=1..m} a_{ix} b_{xj} \]

\[ A \times B = C \quad \text{O(nmk)} \]

**MATRIX-MULTIPLY(A,B)**

1. if columns \([A]\) \# rows \([B]\)
2. then error "incompatible dimensions"
3. else for \( i = 1 \) to rows \([A]\)
4. \( \text{do for } j = 1 \) to columns \([B]\)
5. \( \text{do } C[i,j] = 0 \)
6. \( \text{for } k = 1 \) to columns \([A]\)
7. \( \text{do } C[i,j] = C[i,j] + A[i,k] \times B[k,j] \)
8. return \( C \)

Chain matrix multiplication

- The matrix-chain multiplication problem can be stated as follows: given a chain \( <A_1, A_2, \ldots, A_n> \) of \( n \) matrices
- matrix \( A_i \) has dimension \( p_{i-1} \times p_i \)
- fully parenthesize the product \( A_1 \cdots A_n \) in a way that minimizes the number of scalar multiplications.
\begin{align*}
A_1A_2A_3A_4 \\
& \cdot (A_1(A_2(A_3A_4))) \ , \\
& \cdot (A_1(A_2A_3A_4)) \ , \\
& \cdot ((A_1A_2A_3)A_4) \ , \\
& \cdot (((A_1A_2)A_3)A_4) \ .
\end{align*}

\begin{itemize}
\item Denote the number of alternative parenthesizations of a sequence of \( n \) matrices by \( P(n) \).
\item Since we can split a sequence of \( n \) matrices between the \( k \)th and \( (k + 1) \)st matrices for any \( k = 1, 2, \ldots, n - 1 \) and then parenthesize the two resulting subsequences independently, we obtain the recurrence
\[
P(n) = \begin{cases}
0 & \text{if } n = 1, \\
\sum_{k=1}^{n-1} P(k)P(n-k) & \text{if } n \geq 2.
\end{cases}
\]
\end{itemize}

Let’s crack the problem

\[ A_{i,j} = A_i \bullet A_{i+1} \bullet \cdots \bullet A_j \]

\begin{itemize}
\item Optimal parenthesization of \( A_i \bullet A_{i+1} \bullet \cdots \bullet A_j \) splits at some \( k, k+1 \).
\item Optimal = \( A_{i,k} \bullet A_{k+1,n} \)
\item \( T(A_{i,k}) = T(A_{1,k}) + T(A_{k+1,n}) + T(A_{i,k}) \)
\item \( T(A_{i,k}) \) must be optimal for \( A_i \bullet A_{i+1} \bullet \cdots \bullet A_j \)
\end{itemize}

Recursion

\begin{itemize}
\item \( m[i, j] \) - minimum number of scalar multiplications needed to compute the matrix \( A_{i,j} \)
\item \( m[i,j] = 0 \)
\item \( \text{cost}(A_{i,k} \bullet A_{k+1,j}) = p_{i-1}p_kp_j \)
\item \( m[i, j] = m[i, k] + m[k + 1, j] + p_{i-1}p_kp_j \).
\end{itemize}

This recursive equation assumes that we know the value of \( k \), which we don’t. There are only \( j - i \) possible values for \( k \), however, namely \( k = i, i + 1, \ldots, j - 1 \).

\begin{itemize}
\item Since the optimal parenthesization must use one of these values for \( k \), we need only check them all to find the best. Thus, our recursive definition for the minimum cost of parenthesizing the product \( A_{i,j} \) becomes
\[
m(i,j) = \begin{cases}
0 & \text{if } j = i, \\
\min_{k < j} \{ m[i,k] + m[k+1,j] + p_{i-1}p_kp_j \} & \text{if } j < k.
\end{cases}
\] (16.2)
\item To help us keep track of how to construct an optimal solution, let us define \( s[i, j] \) to be a value of \( k \) at which we can split the product \( A_{i,j} \) to obtain an optimal parenthesization. That is, \( s[i, j] \) equals a value \( k \) such that \( m[i,s[i,j]] + m[s[i,j]+1,j] + p_{i-1}p_kp_j \).
\end{itemize}
Recursion

• Checks all possibilities...

• But – there is only a few subproblems – choose i, j s.t. 1 ≤ i ≤ j ≤ n - O(n^2)

• A recursive algorithm may encounter each subproblem many times in different branches of its recursion tree. This property of overlapping subproblems is the second hallmark of the applicability of dynamic programming.

Example

- A simple inspection of the nested loop structure of MATRIX-CHAIN-ORDER yields a running time of O(n^3) for the algorithm. The loops are nested three deep, and each loop index (l, i, and k) takes on at most n values.
- Time Θ(n^3) ≈ Θ(n^2)
- Space Θ(n^2)

Step 4 of the dynamic-programming paradigm is to construct an optimal solution from computed information.

Use the table s[1..n, 1..n] to determine the best way to multiply the matrices.

Multiply using S table

MATRIX-CHAIN-MULTIPLY(A, s, i, j)
1 if j > i
2 then X = MATRIX-CHAIN-MULTIPLY(A, s, i, s[i, j])
3 Y = MATRIX-CHAIN-MULTIPLY(A, s, s[i, j]+1, j)
4 return MATRIX-MULTIPLY(X, Y)
5 else return A_i

((A_i[A_iA_j]))((A_iA_j)A_j))
Elements of dynamic programming

- **Optimal substructure** within an optimal solution
- **Overlapping subproblems**
- **Memoization**

* A memoized recursive algorithm maintains an entry in a table for the solution to each subproblem. Each table entry initially contains a special value to indicate that the entry has yet to be filled in. When the subproblem is first encountered during the execution of the recursive algorithm, its solution is computed and then stored in the table. Each subsequent time that the subproblem is encountered, the value stored in the table is simply looked up and returned. (tabulated)

- This approach presupposes that the set of all possible subproblem parameters is known and that the relation between table positions and subproblems is established. Another approach is to memoize by using hashing with the subproblem parameters as keys.

**Overlapping subproblems**

**Longest Common Subsequence (LCS)**

**Optimal triangulation**

The problem is to find a triangulation that minimizes the sum of the weights of the triangles in the triangulation

Two ways of triangulating a convex polygon. Every triangulation of this 7-sided polygon has $7 - 3 = 4$ chords and divides the polygon into $7 - 2 = 5$ triangles.

**Parse tree**

Parse trees. (a) The parse tree for the parenthesized product $(A_1 (A_2 A_3)) (A_4 (A_5 A_6))$ and for the triangulation of the 7-sided polygon (b). The triangulation of the polygon with the parse tree overlaid. Each matrix $A_i$ corresponds to the side $v_i v_{i+1}$ for $i = 1, 2, \ldots, 6$. 
Optimal triangulation

\[ d[i,j] = \begin{cases} 0 & \text{if } i = j, \\ \min_{k} \{d[i,k] + d[k+1,j] + w(v_i,v_j)\} & \text{if } i < j. \end{cases} \] (16.7)

Text Algorithms (6EAP)

Similarity measures

Jaak Vilo
2010 fall

Similarity

• How can we measure the similarity of two strings?
• When are the two things “almost” the same?

Edit distance (Levenshtein distance)

• Smallest nr of edit operations to convert one string into the other

How can we calculate this?

• Definition: The edit distance \(D(A,B)\) between strings \(A\) and \(B\) is the minimal number of edit operations to change \(A\) into \(B\). Allowed edit operations are deletion of a single letter, insertion of a letter, or replacing one letter with another.
• Let \(A = a_1 a_2 \ldots a_n\) and \(B = b_1 b_2 \ldots b_m\):
  - E1: Deletion \(a_i \rightarrow \varepsilon\)
  - E2: Insertion \(\varepsilon \rightarrow b_i\)
  - E3: Substitution \(a_i \rightarrow b_i\) (if \(a_i \neq b_i\))
• Other possible variants:
  - E4: Transposition \(a_{i+1} a_{i+2} \rightarrow b_{i+1} b_{i+2}\) and \(a_{i+2} a_{i+1} \rightarrow b_{i+2} b_{i+1}\) (e.g. lecture \rightarrow lecture)
How can we calculate this efficiently?

\[ D(S,T) = \min \begin{cases} 
1. & D(S[1..n-1], T[1..m-1]) + (S[n] \neq T[m]) \cdot 0 : 1 \\
2. & D(S[1..n], T[1..m-1]) + 1 \\
3. & D(S[1..n-1], T[1..m]) + 1 
\end{cases} \]

Define: \( d(i,j) = D(S[1..i], T[1..j]) \)

\[ d(i,j) = \min \begin{cases} 
1. & d(i-1,j-1) + (S[i] \neq T[j]) \cdot 0 : 1 \\
2. & d(i,j-1) + 1 \\
3. & d(i-1,j) + 1 
\end{cases} \]

Recursion?

**Algorithm Edit distance \( D(A,B) \) using Dynamic Programming (DP)**

Input: \( A = a_1a_2...a_n, B = b_1b_2...b_m \)

Output: Value \( d_{mn} \) in matrix \( (d_{ij}) \), \( 0 \leq i \leq m, 0 \leq j \leq n \).

for \( i = 0 \) to \( m \) do \( d_{0i} = i \);

for \( j = 0 \) to \( n \) do \( d_{ij} = j \);

for \( j = 1 \) to \( n \) do

for \( i = 1 \) to \( m \) do

\[ d_{ij} = \min(d_{i-1,j-1} + (\text{if } a_i = b_j \text{ then 0 else 1}), \quad d_{i-1,j} + 1, \quad d_{i,j-1} + 1) \]

return \( d_{mn} \)
21.3.2012

Edit distance is a metric

- It can be shown, that $D(A,B)$ is a metric
  - $D(A,B) \geq 0$, $D(A,B)=0$ iff $A=B$
  - $D(A,B) = D(B,A)$
  - $D(A,C) \leq D(A,B) + D(B,C)$

Alignment

indust-r-y-
in-terest

Path of edit operations

- Optimal solution can be calculated afterwards
  - Quite typical in dynamic programming

Three possible minimizing paths

- Add into $\text{pred}[i,j]$
  - $(i-1,j-1)$ if $d_{ij} = d_{i-1,j-1} + (\text{if } a_i=a_j \text{ then } 0 \text{ else } 1)$
  - $(i-1,j)$ if $d_{ij} = d_{i-1,j} + 1$
  - $(i,j-1)$ if $d_{ij} = d_{i,j-1} + 1$

Three possible minimizing paths

- Add into $\text{pred}[i,j]$
  - $(i-1,j-1)$ if $d_{ij} = d_{i-1,j-1} + (\text{if } a_i=a_j \text{ then } 0 \text{ else } 1)$
  - $(i-1,j)$ if $d_{ij} = d_{i-1,j} + 1$
  - $(i,j-1)$ if $d_{ij} = d_{i,j-1} + 1$

The path (in reverse order) $c \rightarrow c_6, b_5 \rightarrow b_5, c_4 \rightarrow c_4, a_3 \rightarrow a_3, a_2 \rightarrow b_2, a_2 \rightarrow a_2$.
Multiple paths possible

• All paths are correct

• There can be many (how many?) paths

What are the other questions in edit distance calculations?

• Space complexity
• Time complexity
• Other ways to look at the algorithm(s)
• Applications
• More complex notions of similarity
• ...

Space can be reduced

Calculation of $D(A,B)$ in space $\Theta(m)$

Input: $A=a_1a_2...a_m$, $B=b_1b_2...b_n$ (choose $m \leq n$)
Output: $d_{mn}=D(A,B)$
for $i=0$ to $m$ do $C[i]=i$
for $j=1$ to $n$ do

$C=C[0]$;
$C[0]=j$;
for $i=1$ to $m$ do

$d = \min\{ C + (\text{if } a_i=b_j \text{ then } 0 \text{ else } 1), C[i-1]+1, C[i]+1 \}$
$C=C[i]$ // memorize new "diagonal" value
$C[i]=d$
write $C[m]$

Time complexity is $\Theta(mn)$ since $C[0..m]$ is filled $n$ times

Shortest path in the graph


All nodes at distance 1 from source
21.3.2012

Observations?

- Shortest path is close to the diagonal
  - If a short distance path exists
- Values along any diagonal can only increase (by at most 1)

Transform the matrix into $f_{kp}$

- For each diagonal only show the position (row index) where the value is increased by 1.
- Also, one can restrict the matrix $(d_{ij})$ to only this part where $d_{ij} \leq d_{\text{max}}$ since only those $d_{ij}$ can be on the shortest path.
- We'll use the matrix $(f_{kp})$ that represents the diagonals of $d_{ij}$
  - $f_{kp}$ is a row index $i$ from $d_{ij}$ such that on diagonal $k$ the value $p$ reaches row $i$ ($d_{ip}=p$ and $j-i=k$).
  - Initialization: $f_{i-1} = 1$ and $f_{p} = \infty$ when $p \leq |k|-1$;
  - $d_{\text{max}} = p$, such that $f_{1 \text{max}} = m$

Diagonal

$$
\begin{array}{cccccc}
  & a & b & c & b & c \\
\hline
 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\
 2 & 1 & 1 & 2 & 3 & 4 \\
 3 & 2 & 1 & 2 & 3 & 4 \\
 4 & 3 & 3 & 3 & 3 & 3 \\
 5 & 4 & 4 & 4 & 4 & 4 \\
\end{array}
$$

Diagonal lemma

Lemma: For each $d_{ij}$, 1st item, 1st item holds: $d_{ij} = d_{i-1,j-1}$ or $d_{ij} = d_{i,j-1} + 1$.

(notice that $d_{i-1,j-1}$ and $d_{i,j-1}$ are on the same diagonal)

Proof: Since $d_{ij}$ is an integer, show:
- $d_{ij} \leq d_{i-1,j-1} + 1$
- $d_{ij} \geq d_{i-1,j-1}$

From the definition of edit distance 1. holds since $d_{ij} \leq d_{i-1,j-1} + 1$

Induction on $i+j$:
- Basis is trivial when $i=0$ or $j=0$ (if we agree that $d_{-1,j} = d_{i-1}$)
- Induction step: there are 3 possibilities:
  - On minimization the $d_{ij}$ is calculated from entries $d_{i-1,j-1}$ and $d_{i-1,j}$
  - On minimization the $d_{ij}$ is calculated from entries $d_{i,j-1}$ and $d_{i,j}$
  - On minimization the $d_{ij}$ is calculated from entry $d_{i-1,j}$. Analogical to 2.
- Hence, $d_{i-1,j} \leq d_{ij}$

Diagonal nr. 2, $d_{22}, d_{23}, d_{24}, d_{25}, d_{26}$, $d_{02}, d_{13}, d_{24}, d_{35}, d_{46}$

$$
\begin{array}{cccccc}
  & a & b & c & b & c \\
\hline
 1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\
 2 & 1 & 1 & 2 & 3 & 4 \\
 3 & 2 & 1 & 2 & 3 & 4 \\
 4 & 3 & 3 & 3 & 3 & 3 \\
 5 & 4 & 4 & 4 & 4 & 4 \\
\end{array}
$$

Transform the matrix into $f_{kp}$

- For each diagonal only show the position (row index) where the value is increased by 1.
- Also, one can restrict the matrix $(d_{ij})$ to only this part where $d_{ij} \leq d_{\text{max}}$ since only those $d_{ij}$ can be on the shortest path.
- We’ll use the matrix $(f_{kp})$ that represents the diagonals of $d_{ij}$
  - $f_{kp}$ is a row index $i$ from $d_{ij}$ such that on diagonal $k$ the value $p$ reaches row $i$ ($d_{ip}=p$ and $j-i=k$).
  - Initialization: $f_{i-1} = 1$ and $f_{p} = \infty$ when $p \leq |k|-1$;
  - $d_{\text{max}} = p$, such that $f_{1 \text{max}} = m$

Diagonal lemma

Lemma: For each $d_{ij}$, 1st item, 1st item holds: $d_{ij} = d_{i-1,j-1}$ or $d_{ij} = d_{i,j-1} + 1$.

(notice that $d_{i-1,j-1}$ and $d_{i,j-1}$ are on the same diagonal)

Proof: Since $d_{ij}$ is an integer, show:
- $d_{ij} \leq d_{i-1,j-1} + 1$
- $d_{ij} \geq d_{i-1,j-1}$

From the definition of edit distance 1. holds since $d_{ij} \leq d_{i-1,j-1} + 1$

Induction on $i+j$:
- Basis is trivial when $i=0$ or $j=0$ (if we agree that $d_{-1,j} = d_{i-1}$)
- Induction step: there are 3 possibilities:
  - On minimization the $d_{ij}$ is calculated from entries $d_{i-1,j-1}$ and $d_{i-1,j}$
  - On minimization the $d_{ij}$ is calculated from entries $d_{i,j-1}$ and $d_{i,j}$
  - On minimization the $d_{ij}$ is calculated from entry $d_{i-1,j}$. Analogical to 2.
- Hence, $d_{i-1,j} \leq d_{ij}$
Calculating matrix \( (f_{kp}) \) by columns

- Assume the column \( p-1 \) has been calculated in \( (f_{kp}) \), and we want to calculate \( f_{kp} \) (the region of \( d_{ij} = p \)).
- On diagonal \( k \) values \( p \) reach at least the row \( t = \max \{ f_{k,p-1}+1, f_{k-1,p-1}, f_{k+1,p-1}+1 \} \) if the diagonal \( k \) reaches so far.
- If on row \( t+1 \) additionally \( a_i = b_j \) on the same diagonal, then \( d_{ij} \) cannot increase, and value \( p \) reaches row \( t+1 \).
- Repeat previous step until \( a_i \neq b_j \) on diagonal \( k \).

Algorithm A(): calculate \( f_{kp} \)

\[
\begin{align*}
A(k,p) & \quad t = \max(f_{k,p-1}+1, f_{k-1,p-1}, f_{k+1,p-1}+1) \\
& \quad \text{while } a_{t+1} = b_{t+1+k} \text{ do } t = t+1 \\
& \quad f_{kp} = \begin{cases} 
\text{undefined} & \text{if } t > m \text{ or } t+k > n \\
\text{else } t & \text{if } 0, 2, \ldots \text{t} 
\end{cases}
\end{align*}
\]

Algorithm: Diagonal method by columns

\[
\begin{align*}
p & = -1 \\
& \text{while } f_{n-m,p} \neq m \\
& \quad p = p+1 \\
& \quad \text{for } k = -p \text{ to } p \text{ do } / / \quad f_{kp} = A(k,p) \\
& \quad \quad t = \max(f_{k,p-1}+1, f_{k-1,p-1}, f_{k+1,p-1}+1) \\
& \quad \quad \text{while } a_{t+1} = b_{t+1+k} \text{ do } t = t+1 \\
& \quad \quad f_{kp} = \begin{cases} 
\text{undefined} & \text{if } t > m \text{ or } t+k > n \\
\text{else } t & \text{if } 0, 2, \ldots \text{t} 
\end{cases}
\end{align*}
\]

- \( p \) can only occur on diagonals \(-p \leq k \leq p\).
- Method can be improved since \( k \) is often such that \( f_{kp} \) is undefined.
- We can decrease values of \( k \):
  - \(-m \leq k \leq n\) (diagonal numbers)
  - Let \( m \leq n \) and \( d_{ij} \) on diagonal \( k \):
    - if \(-m \leq k \leq 0 \) then \( |k| \leq d_i \leq m \)
    - if \( 1 \leq k \leq n \) then \( k \leq d_i \leq k+m \)
    - Hence, \(-m \leq k \leq m\) if \( p \leq m \) and \( p-m \leq k \leq p \) if \( p \geq m \)
Some notes

- In applications small $D(A,B)$ are most interesting.
- Can modify the algorithm by providing maximum $t$
- Hence, $O(tm)$ - the smaller the $t$, the faster the algorithm.
- Space can be reduced by keeping only previous column
- How to output the shortest path?
  - Relatively simple, in time $O(s)$, outputs a single path.

Extensions to basic edit distance

- New operations
- Variable costs
- ...

Transposition ($ab \rightarrow ba$)

- **E4: Transposition**
  
  $a_{i+1} \rightarrow b_{j+1}$, s.t. $a_i = b_j$ and $a_{i+1} = b_j$

  - (e.g.: lecture $\rightarrow$ letcure)

  

<table>
<thead>
<tr>
<th>$d(i,j)$</th>
<th>$d(i-1,j-1) + (S[n]=T[m]) ? 0 : 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d(i,j)$</td>
<td>$d(i,j-1) + 1$</td>
</tr>
<tr>
<td>$d(i,j)$</td>
<td>$d(i-1,j) + 1$</td>
</tr>
<tr>
<td>$d(i,j)$</td>
<td>$d(i-2,j-2) + (\text{if } S[i-1,i] = T[j,j-1] \text{ then } 1 \text{ else } \infty)$</td>
</tr>
</tbody>
</table>

Longest common subsequences

- **Definition.** String $C = c_1 c_2 ... c_r$ is a subsequence (alamjada) of $A = a_1 a_2 ... a_m$ if by removing from $A$ null or more characters, one can get $C$.
- String $C = c_1 c_2 ... c_r$ is the **longest common subsequence**, LCS (pikim uhine alamjada) of $A = a_1 a_2 ... a_m$ and $B = b_1 b_2 ... b_n$, if $C$ is the longest string that is both the subsequence of $A$ and $B$.
- $C = \text{LCS}(A,B)$
- The length of $\text{LCS}(A,B)$, $|C|$, can be used as the similarity measure for $A$ and $B$.
- $\text{LCS}(A,B)$ can be calculated similarly to edit distance
\[|LCS(A,B)| = (|A| + |B| - D'(A,B))/2\]

- Let \(D'(A,B)\) the edit distance where the \textit{only allowed operations} are insertion and deletion (no replace).

- \textbf{Theorem}
  
  \textbf{a)} \(|LCS(A,B)| = (|A| + |B| - D'(A,B))/2\)
  
  \textbf{b)} Let have two sets \(D'(A,B)\) with optimal nr of changes:
  
  1. \(x_i \rightarrow \varepsilon, x_{i+1} \rightarrow \varepsilon, \ldots, x_p \rightarrow \varepsilon\) deletions from A and
  
  2. \(\varepsilon \rightarrow y_j, \varepsilon \rightarrow y_{j+1}, \ldots, \varepsilon \rightarrow y_r\) insertions into B.
  
  Then \(LCS(A,B)=C\) can be constructed such that, C is A after deletions of 1. and C is B after deletion of all insertions 2. (insertions in 2. are reversely deletions from B).

\textbf{Proof b)}

- According to construction, \(C\) is uniquely defined
  
  \(C\) is a subsequence of A as well as B.
  
  If \(C\) was not the longest, then there would be\(C', |C| < |C'|\) s.t. \(C'=LCS(A,B)\).

- But then \(D'(A,B)=|A|+|B|-|C'| < |A|-|C|+|B|-|C| = D'(A,B)\), which is a contradiction.

- Hence, \(C=LCS(A,B)\).

\textbf{Proof a)}

- According to \(b)\) \(|LCS(A,B)|=|A|-p\) and \(|LCS(A,B)|=|B|-r\), or 2 |

\(|LCS(A,B)|=|A|+|B|-|p+r|=|A|+|B|-D'(A,B)|.

- \textbf{Example}. \(LCS(\text{england, inglismaa})=\text{ngla}.

\(D'(\text{england, inglismaa})=8, \ |\text{ngla}|=4=(7+9-8)/2.\)

- Diagonal lemma holds, but the increase always occurs by two.

- Time complexity \(O(mn)\), with diag. method \(O\ (\min(s,m)\ s)\) where \(s=D'(A,B), m=|A|, n=|B|\).

- There exists other algorithms for LCS (e.g. Hunt-Szymanski)

- Unix command \texttt{diff} compares files row by row and searches the deviations from the LCS of the two files.

\textbf{Generalized edit distance}

- Use more operations \(E_1...E_n\), and to provide different costs to each.

- \textbf{Definition}. Let \(x, y \in \Sigma^*\). Then every \(x \rightarrow y\) is an edit operation. Edit operation replaces \(x\) by \(y\).

  - if \(Axyuv\) then after the operation, \(Axyuv\)

  - We note by \(w(x \rightarrow y)\) the cost or weight of the operation.

  - Cost may depend on \(x\) and/or \(y\). But we assume \(w(x \rightarrow y) \geq 0.\)

- \textbf{Generalized edit distance}

  - If operations can only be applied in parallel, i.e. the part already changed cannot be modified again, then we can use the dynamic programming.

  - Otherwise it is an algorithmically unsolvable problem, since question - can \(A\) be transformed into \(B\) using operations of \(G\), is unsolvable.

  - The diagonal method in general may not be applicable.

  - But, since each diversion from diagonal, the cost slightly increases, one can stay within the narrow region around the diagonal.
Applications of generalized edit distance

- Historic documents, names
- Human language and dialects
- Transliteration rules from one alphabet to another
e.g. Tõugu => Tyugu (via Russian)
- ...

Examples

Examples of applications:

- Historic documents, names
- Human language and dialects
- Transliteration rules from one alphabet to another
e.g. Tõugu => Tyugu (via Russian)
- ...

How?

- Apply Aho-Corasick to match for all possible edit operations
- Use minimum over all possible such operations and costs
- Implementation: Reina Käärik, Siim Orasmaa
Possible problems/tasks

- Manually create sensible lists of operations
  - For English, Russian, etc...
  - Old language,
- Improve the speed of the algorithm (testing)
- Train for automatic extraction of edit operations and respective costs from examples of matching words...

Advanced Dynamic Programming

- Robert Giegerich:
  - [http://www.techfak.uni-bielefeld.de/ags/pi/lehre/ADP/](http://www.techfak.uni-bielefeld.de/ags/pi/lehre/ADP/)
- Algebraic dynamic programming
  - Functional style
  - Haskell compiles into C