CMPT 413: Computational Linguistics
ED3: Edit Distance and FSTs

Anoop Sarkar

http://www.cs.sfu.ca/~anoop

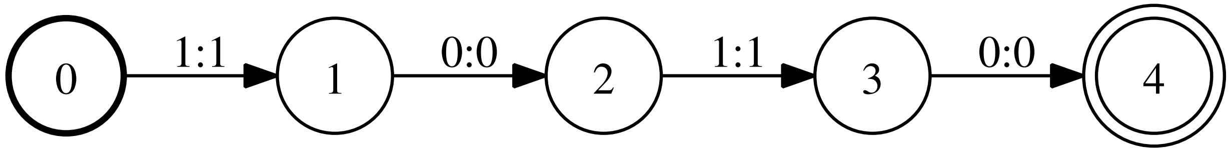
Edit Distance and FSTs

- Algorithm using a Finite-state transducer:
  - Construct an Edit FST:
    - a transition x:x gets zero cost for every x in the alphabet
    - a transition on ε:x (insertion) gets cost 1
    - x:ε (deletion) for any char x gets cost 1
    - optionally x:y (substitution) gets a cost 1 for every x, y
  - Compose source FST, edit FST and target FST
  - Finding minimum cost edit distance == Finding the shortest path from start state to final state
  - Computes an error rate
    - For words this computes the word error rate
    - For letters the letter error rate

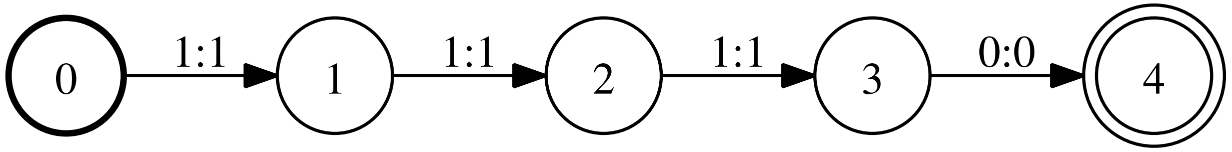
Edit Distance and FSTs

- Lets assume we want to edit source string 1010 into the target string 1110
- The alphabet is just 1 and 0

SOURCE

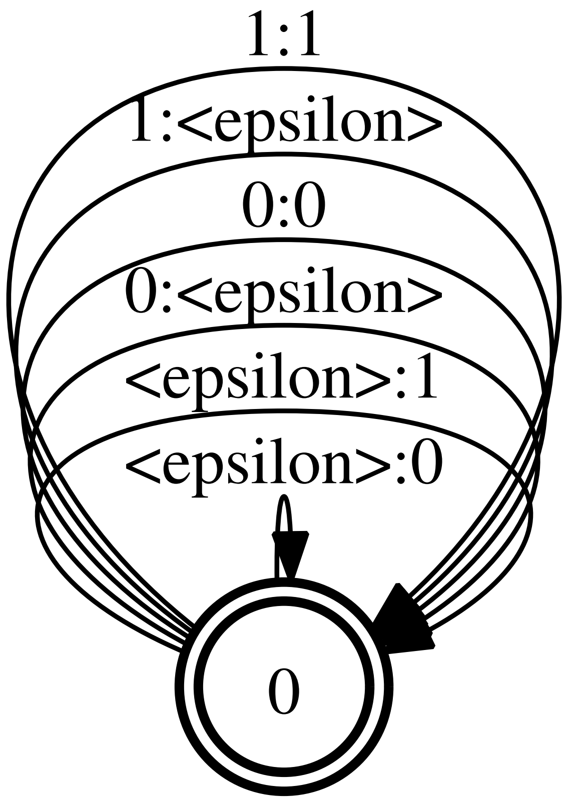


TARGET



Edit Distance and FSTs

- Construct a FST that allows strings to be edited: aka flower FST

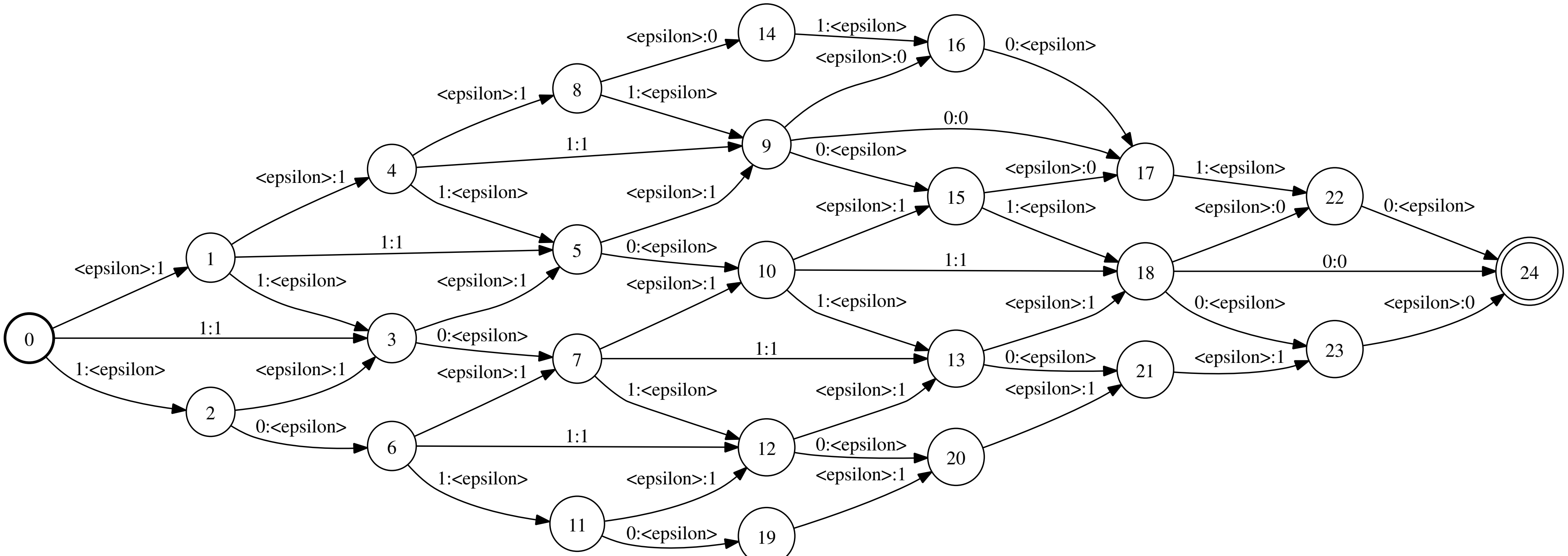


EDITS

Subst cost = 2

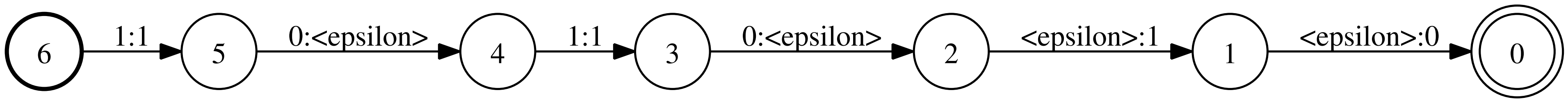
Edit Distance and FSTs

- Compose SOURCE and EDITS and TARGET



Edit Distance and FSTs

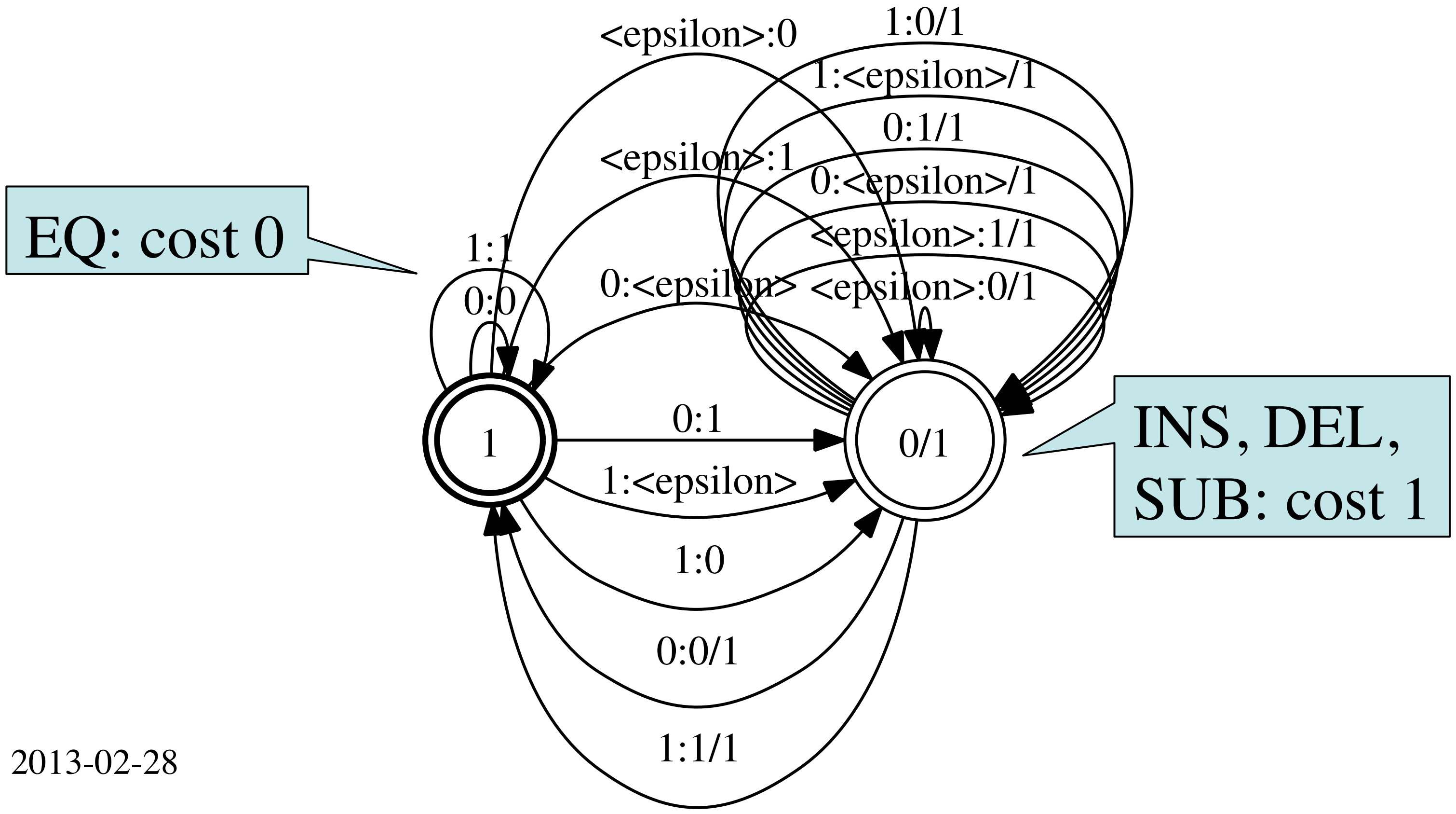
- The shortest path is the minimum edit FST from SOURCE (1010) to TARGET (1110)



$$\begin{tikzcd}
6 \arrow[r, "1:1"] & 5 \arrow[r, "0:\text{<epsilon>}"] & 4 \arrow[r, "1:1"] & 3 \arrow[r, "0:\text{<epsilon>}"] & 2 \arrow[r, "\text{<epsilon>}:1"] & 1 \arrow[r, "\text{<epsilon>}:0"] & 0
\end{tikzcd}$$

Edit Distance and FSTs

However, if we want a substitution cost of 1 (instead of 2) then we have to create a different transducer.

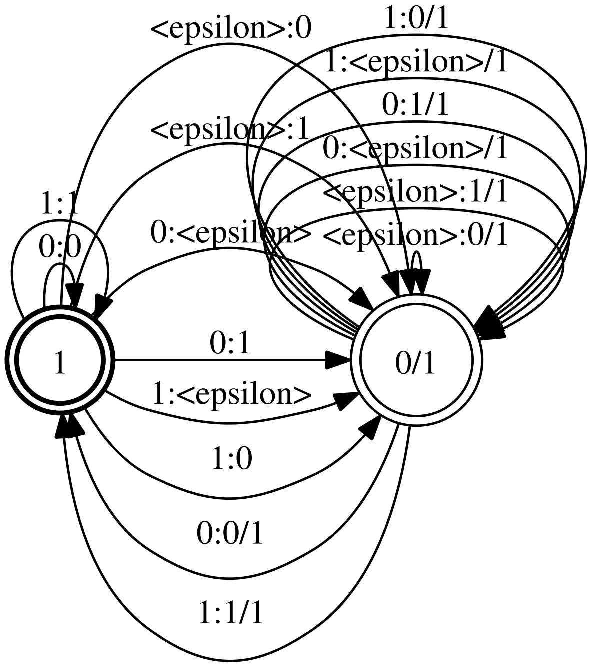


Edit distance and FSTs

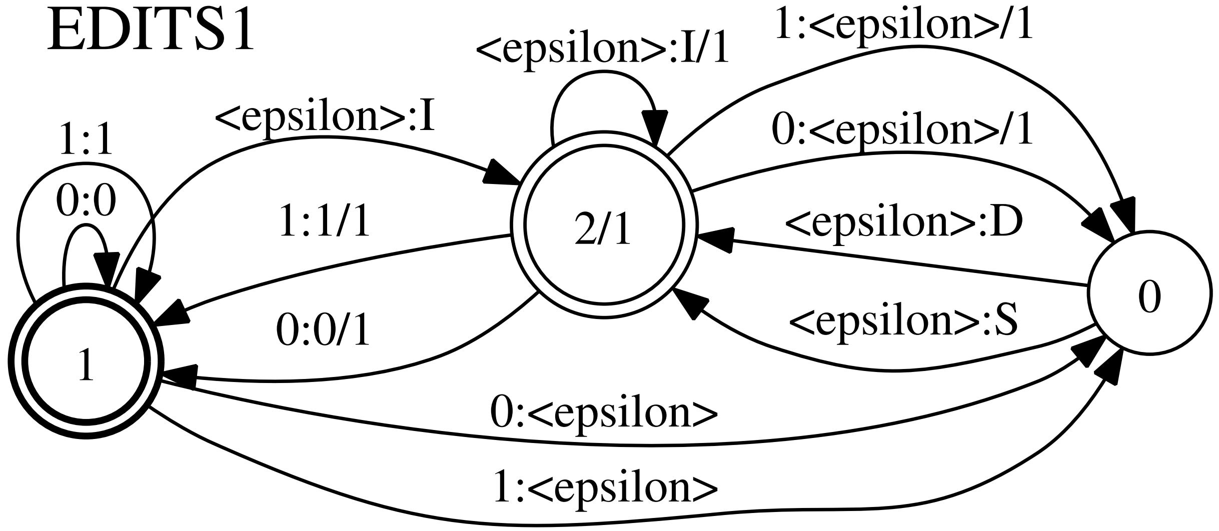
- One problem is scaling to larger character sets
- For 95 ascii symbols the Levenshtein edit transducer will have 9215 transitions
- For 10,000 words the Edit FST needs 100,020,000 (100M) transitions
- Number of transitions = $(V+1)^2$-1
- Solution: De-compose the Edit FST

has $(V+1)^2-1$ transitions

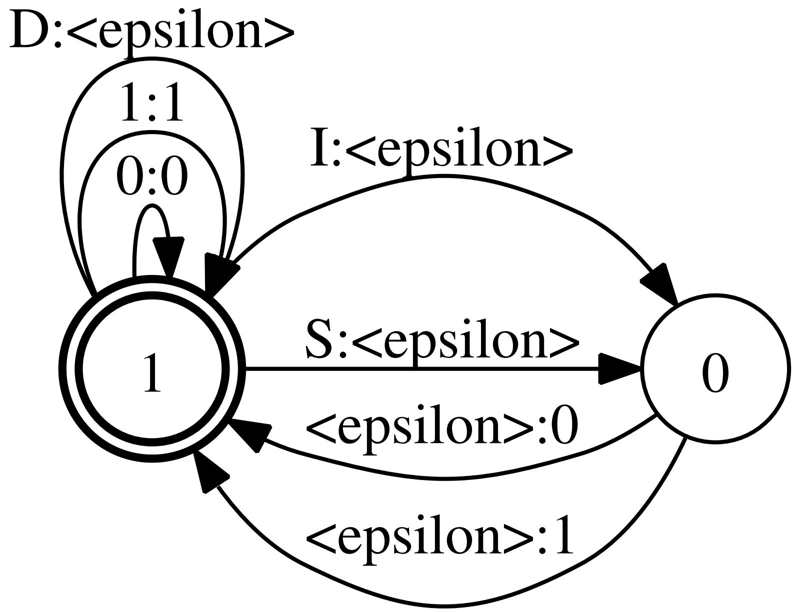
EDITS =
EDITS1 ° EDITS2



EDITS1



has ≈4V transitions



EDITS2

has ≈4V transitions

CMPT 413: Computational Linguistics
ED3: Edit Distance and FSTs

Anoop Sarkar

http://www.cs.sfu.ca/~anoop