Sequence to sequence models

11-785 Introduction to Deep Learning

- lecture 16 -

TAVE Research DL001

Heeji Won

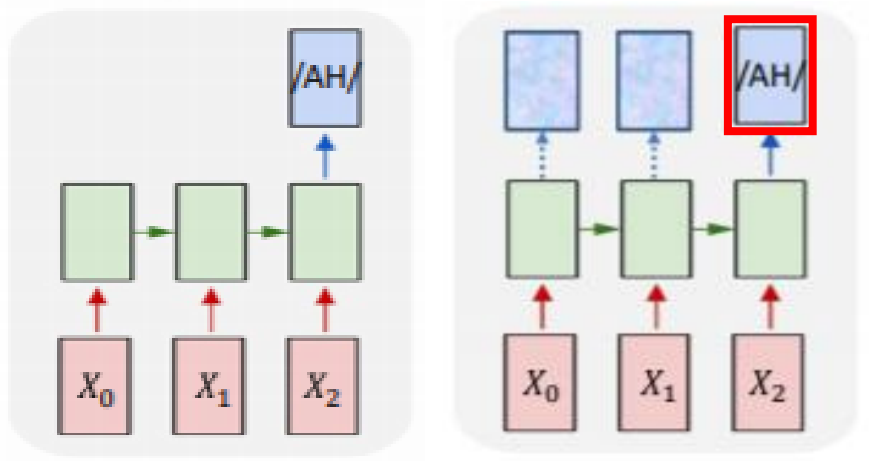
Contents

1. Many to one model
2. Sequence to sequence model

Contents

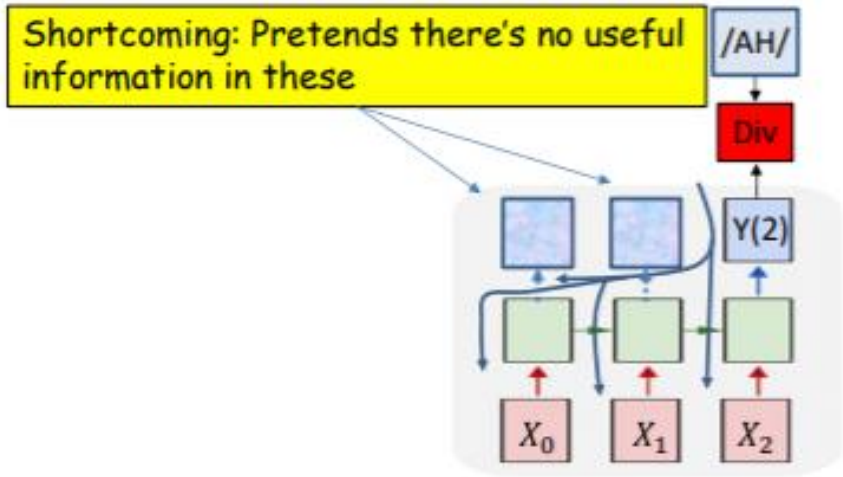
01. Many to one model

- Many to one
  - used for Q&A, Speech recognition

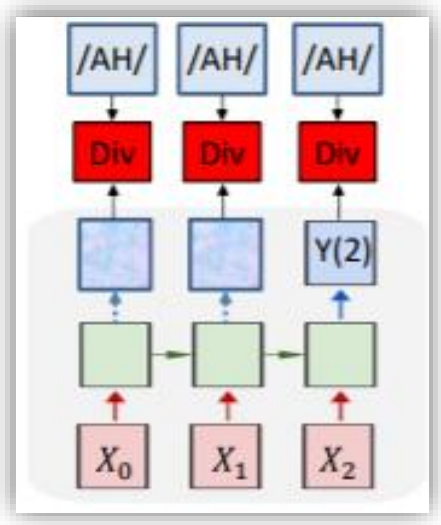


  - Outputs are actually produced for every input
  - But, we only read it at the end of the sequence

- Training



=> Exploit them! Assume the same output for the entire input



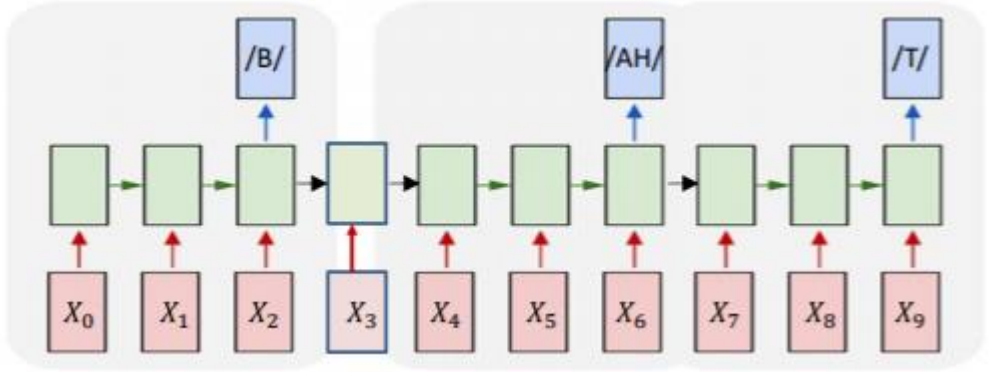
Define the divergence everywhere!

$$DIV(Y_{target}, Y) = \sum_t w_t Xent(Y(t), Phoneme)$$

Contents

02. Sequence to sequence model

- Sequence to sequence
  - Order synchronous, but time asynchronous
  - E.g. phoneme recognition, speech recognition

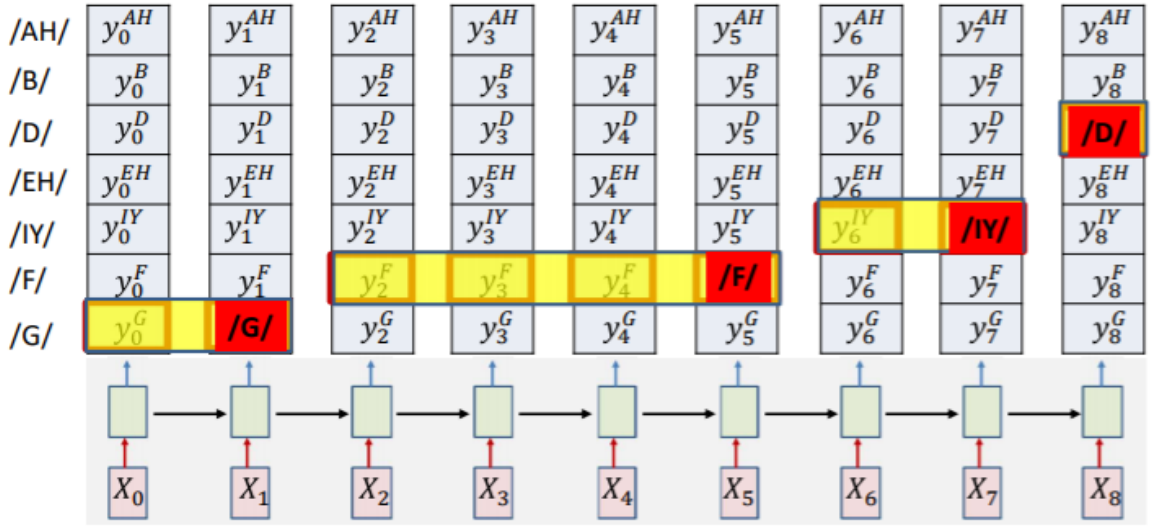


✓ How do we know when to output symbols?

- In fact, the network produces outputs at every time

➢ Where to output

- Option 1 : Simply select the most probable symbol at each time

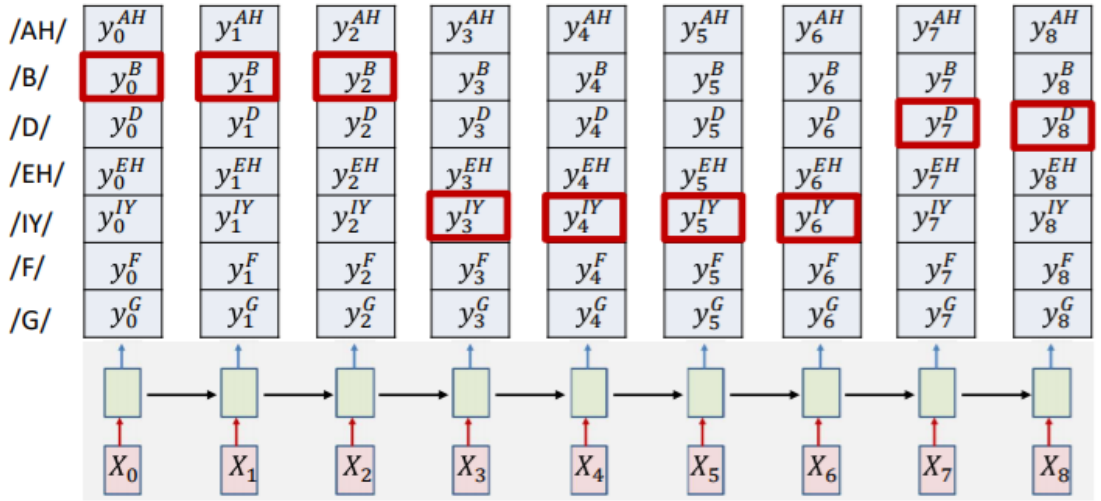


- Merge adjacent repeated symbols and place the actual emission of the symbol in the final instant
- But, resulting sequence may be meaningless

02. Sequence to sequence model

➢ Where to output

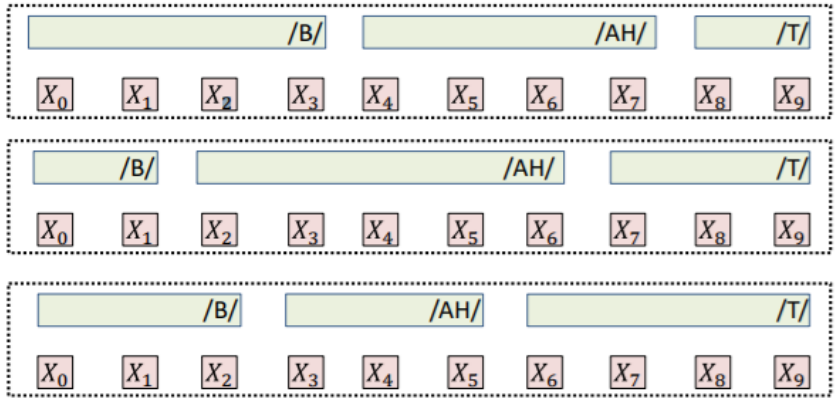
- Option 2 : Simply select the most probable symbol at each time



- E.g. only allow sequences corresponding to dictionary words
- This is a suboptimal decode that finds the most likely time-synchronous output sequence

➢ Training

- There can be various alignment of labels

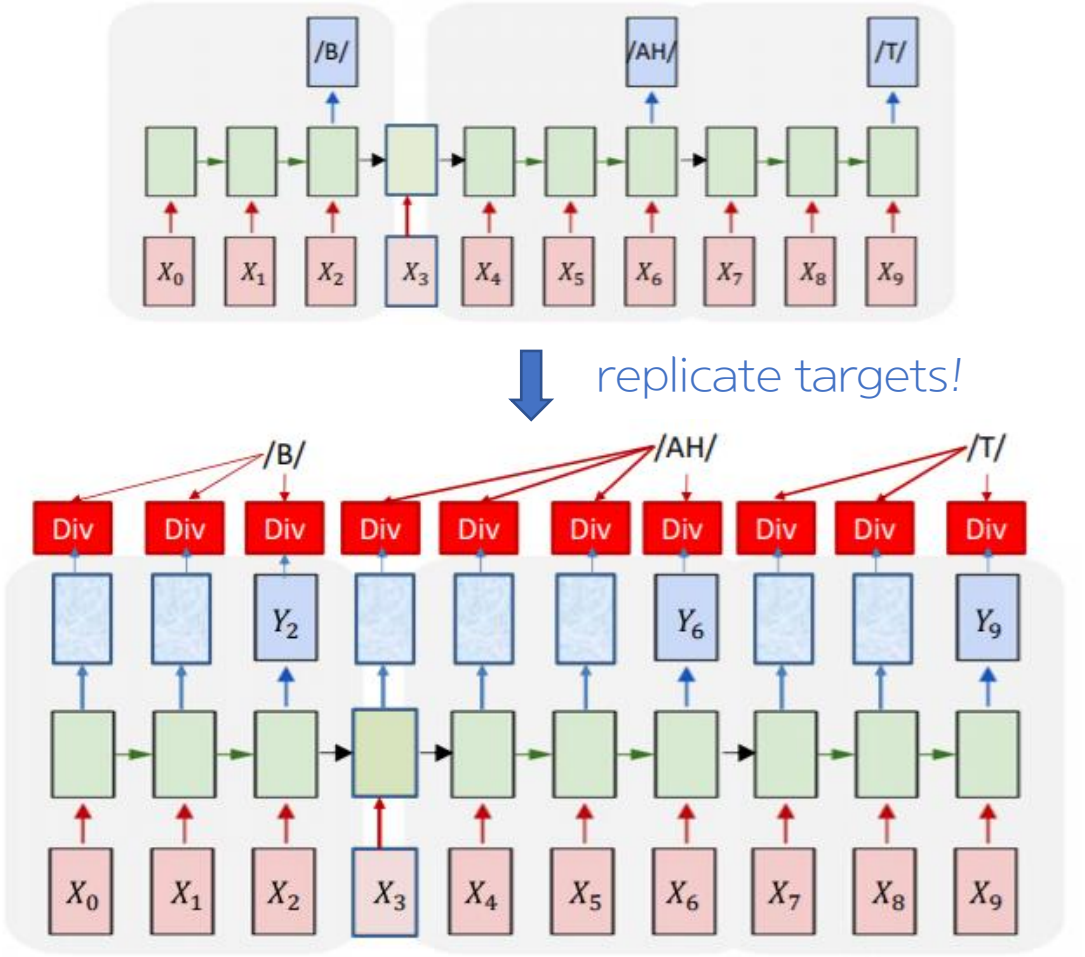


Situation 1. When we know the alignment

Situation 2. When we don' t know the alignment

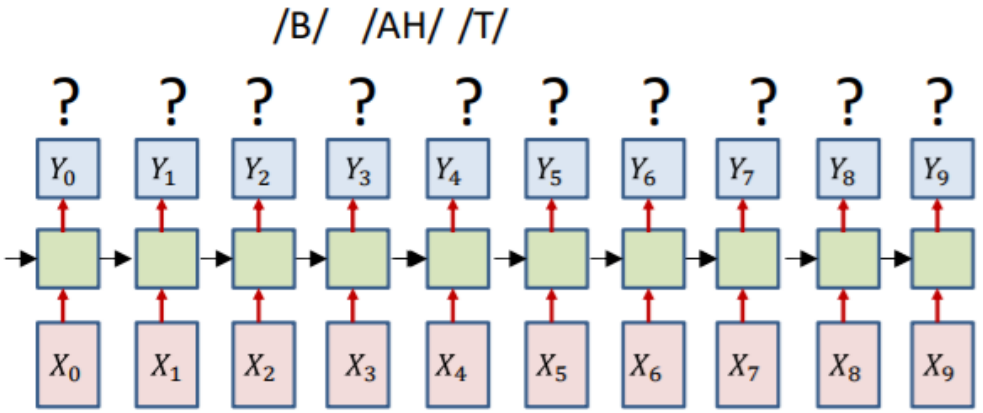
02. Sequence to sequence model

- Training with alignment



Time synchronous expansion of order synchronous seq.

- Training without alignment



Situation 1. Guess the alignment

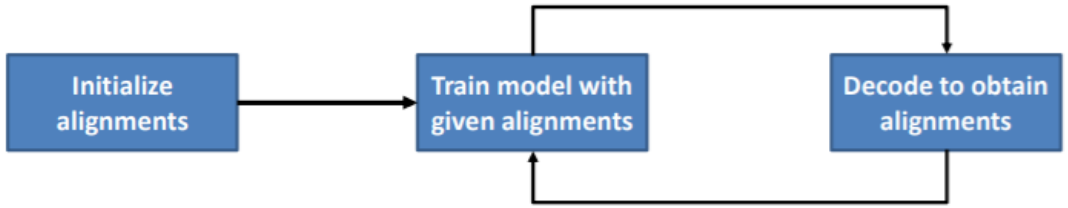
Situation 2. Consider all possible alignment

02. Sequence to sequence model

- Solution 1 : Guess the alignment

– Guess an initial alignment and iteratively refine as the model improves

– Initialize either randomly, based on some heuristic, or any other rationale



– How to estimate an alignment

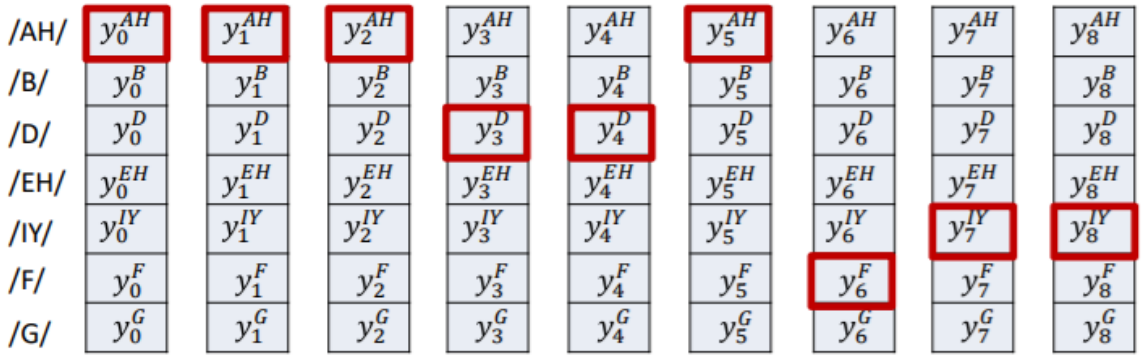
Find

$$argmax\ P(s_0, s_1, \dots, s_{N-1} | S_0, S_1, \dots, S_K, X_0, X_1, \dots, X_{N-1})$$

$$compress(s_0, s_1, \dots, s_{N-1}) \equiv S_0, S_1, \dots, S_K$$

➢ Decoding

- Unconstrained decoding

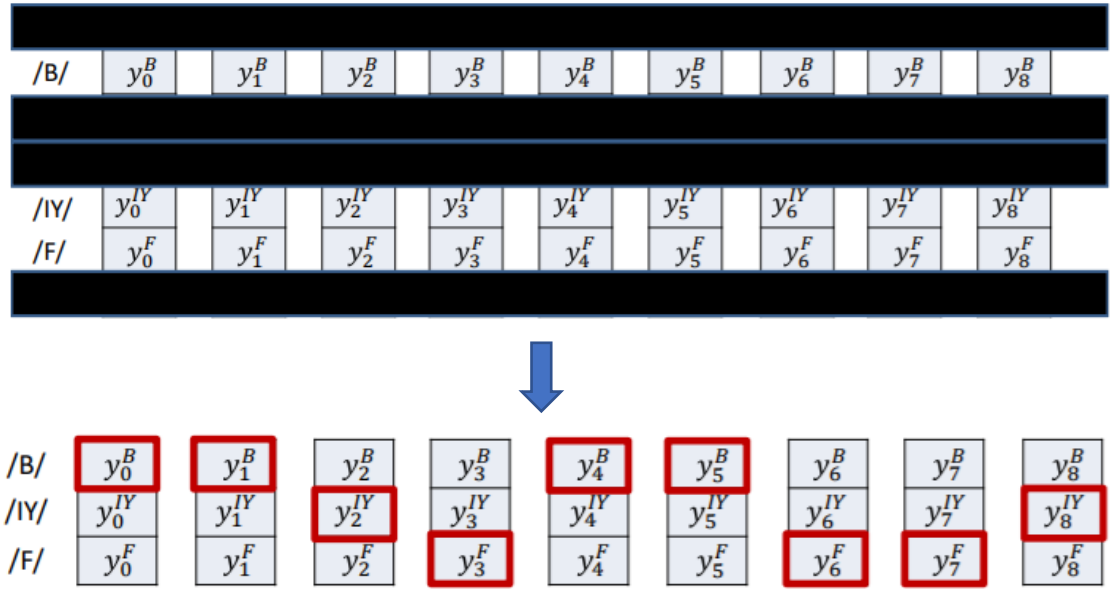


– target : /B/ /IY/ /F/ /IY/,
– output : /AH/ /AH/ ⋯ /IY/

✓ Output may not correspond to an expansion of the desired symbol seq.

02. Sequence to sequence model

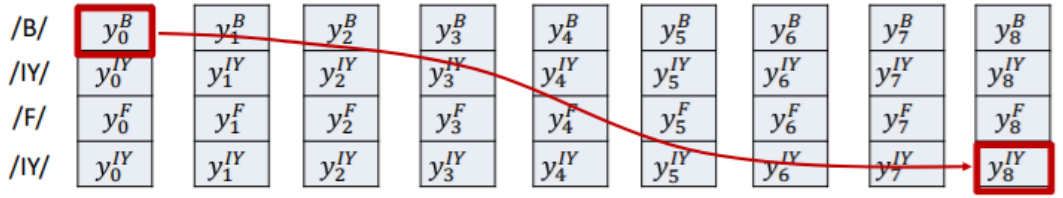
- Block out
  - Block out all rows that do not include symbols from the target sequence



  - Only decode on reduced grid

✓ Still not assure that the decode sequence expands the target symbol seq.

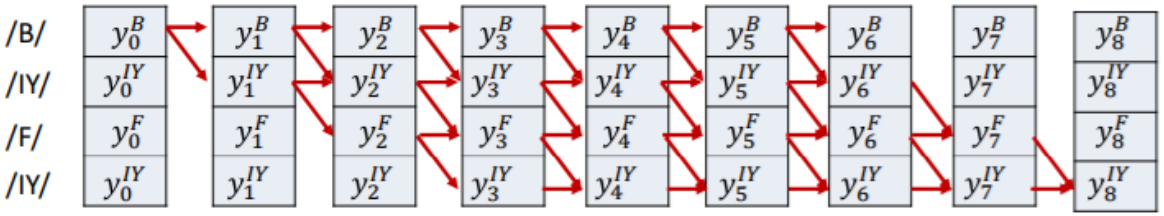
- Explicitly constrain alignment
  - Arrange the constructed table which has exact target sequence
  - the first symbol must be the top left block and the last symbol must be the bottom right
  - The rest of symbols must monotonically travel down form top left to bottom right



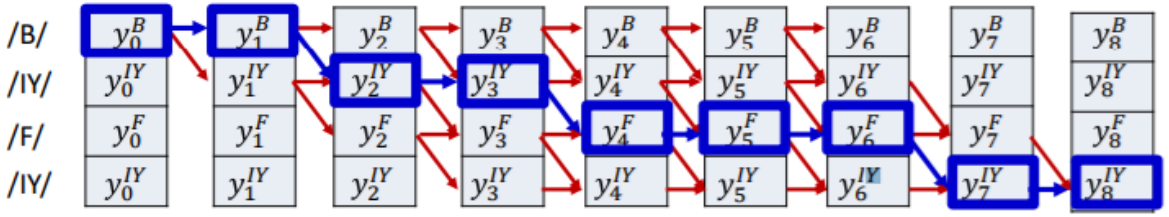
✓ The seq. is an expansion of the target seq.

02. Sequence to sequence model

- The graph representing all path

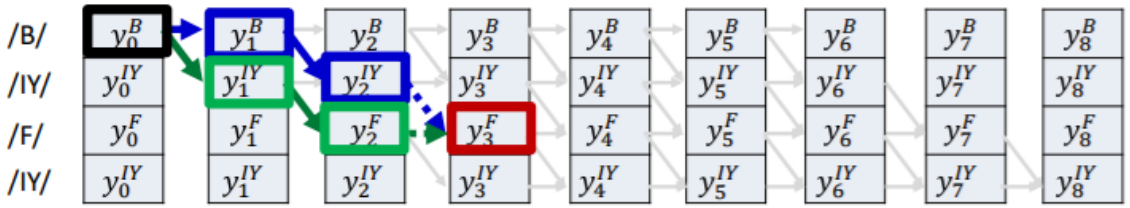


  - Find the most probable path using any dynamic programming algorithm like the **Viterbi algorithm**



➢ Verterbi algorithm

  - The best path to any node must be an extension of the best path to one of its parent nodes
  - Dynamically track the best path



  - Initialization

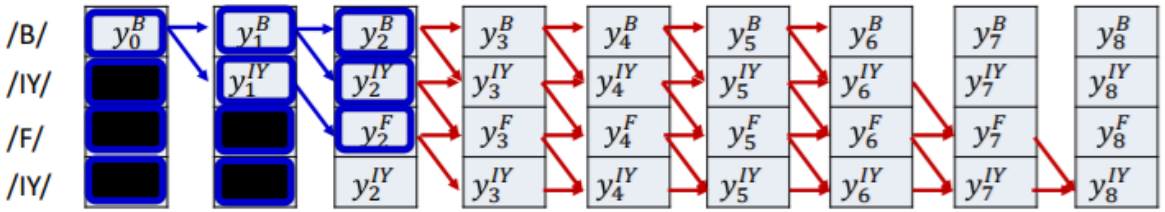
BP := Best Parent
Bscr := Bestpath Score to node

$$BP(0,i) = null,\ i = 0 \ldots K-1$$

$$Bscr(0,0) = y_0^{S(0)},\ Bscr(0,i) = -\infty,\ i = 1 \ldots K-1$$

02. Sequence to sequence model

➢ Verterbi algorithm



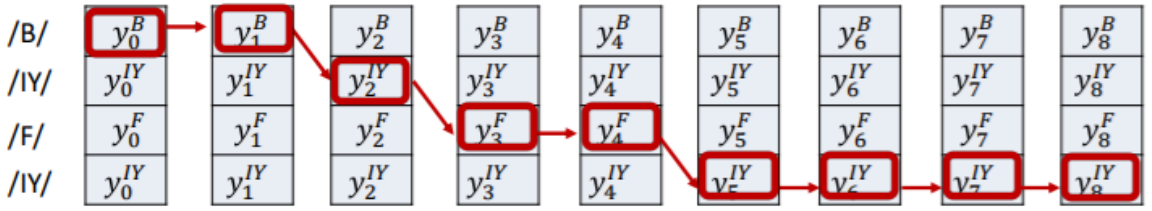
for $t = 1 \ldots T-1$

$BP(t,0) = 0;\ Bscr(t,0) = Bscr(t-1,0) \times y_t^{S(0)}$

for $l = 1 \ldots K-1$

- $BP(t,l) = \left(if\ \left(Bscr(t-1,l-1) > Bscr(t-1,l)\right)\ l-1;\ else\ l\right)$
- $Bscr(t,l) = Bscr(BP(t,l)) \times y_t^{S(l)}$

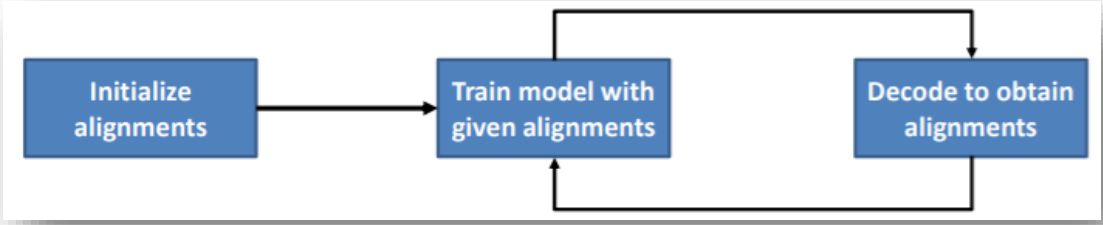
- Gradients from the alignment



$$DIV = \sum_t KL\left(Y_t, symbol_t^{bestpath}\right) = -\sum_t \log Y\left(t, symbol_t^{bestpath}\right)$$

$$\nabla_{Y_t} DIV = \left[0 \quad 0 \quad \cdots \quad \frac{-1}{Y\left(t, symbol_t^{bestpath}\right)} \quad 0 \quad \ldots \quad 0\right]$$

– The gradient is 0 except the component corresponding to the target (estimated alignment)



Thank you