Sequence to Sequence Models

Deep Learning

Woohwan Jung

한양대학교
HANYANG UNIVERSITY

DSLAB
데이터사이언스 연구실

Basic Sequence2Sequence Model

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$

Jane visite l'Afrique en septembre

(문맥 벡터 만들기

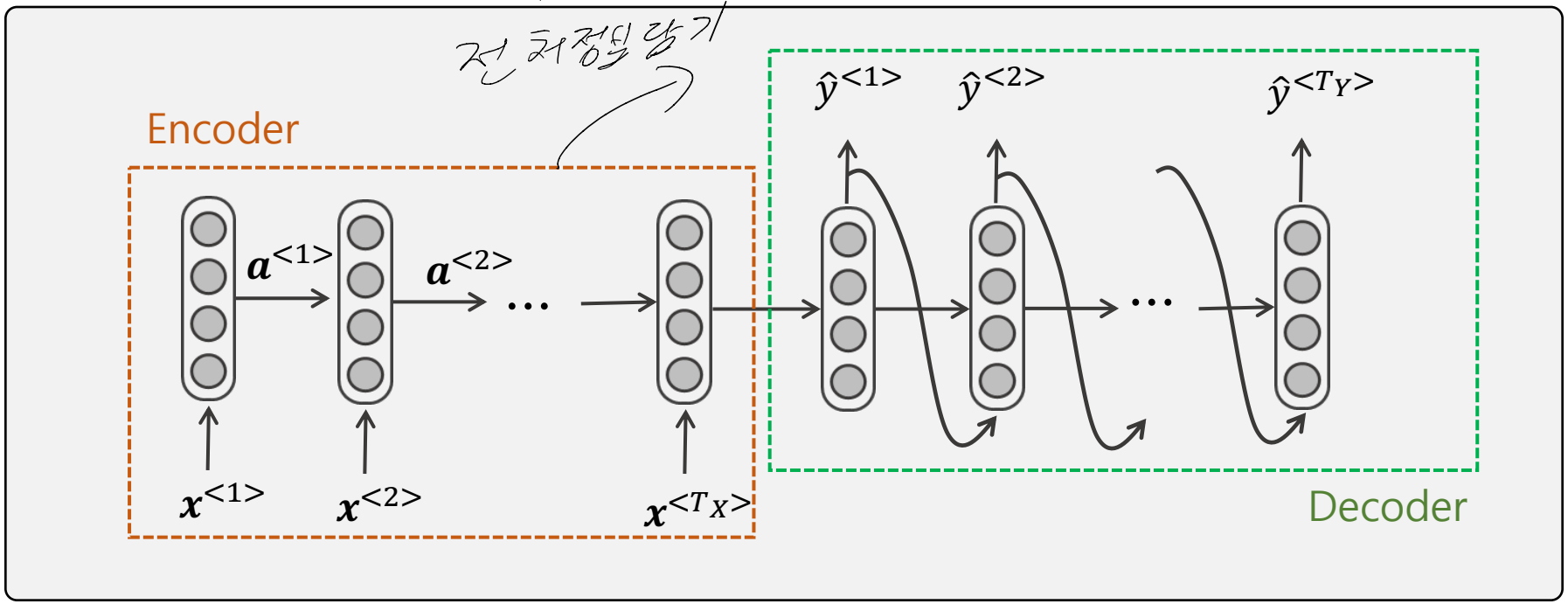
Attention model/사용

번역을 위

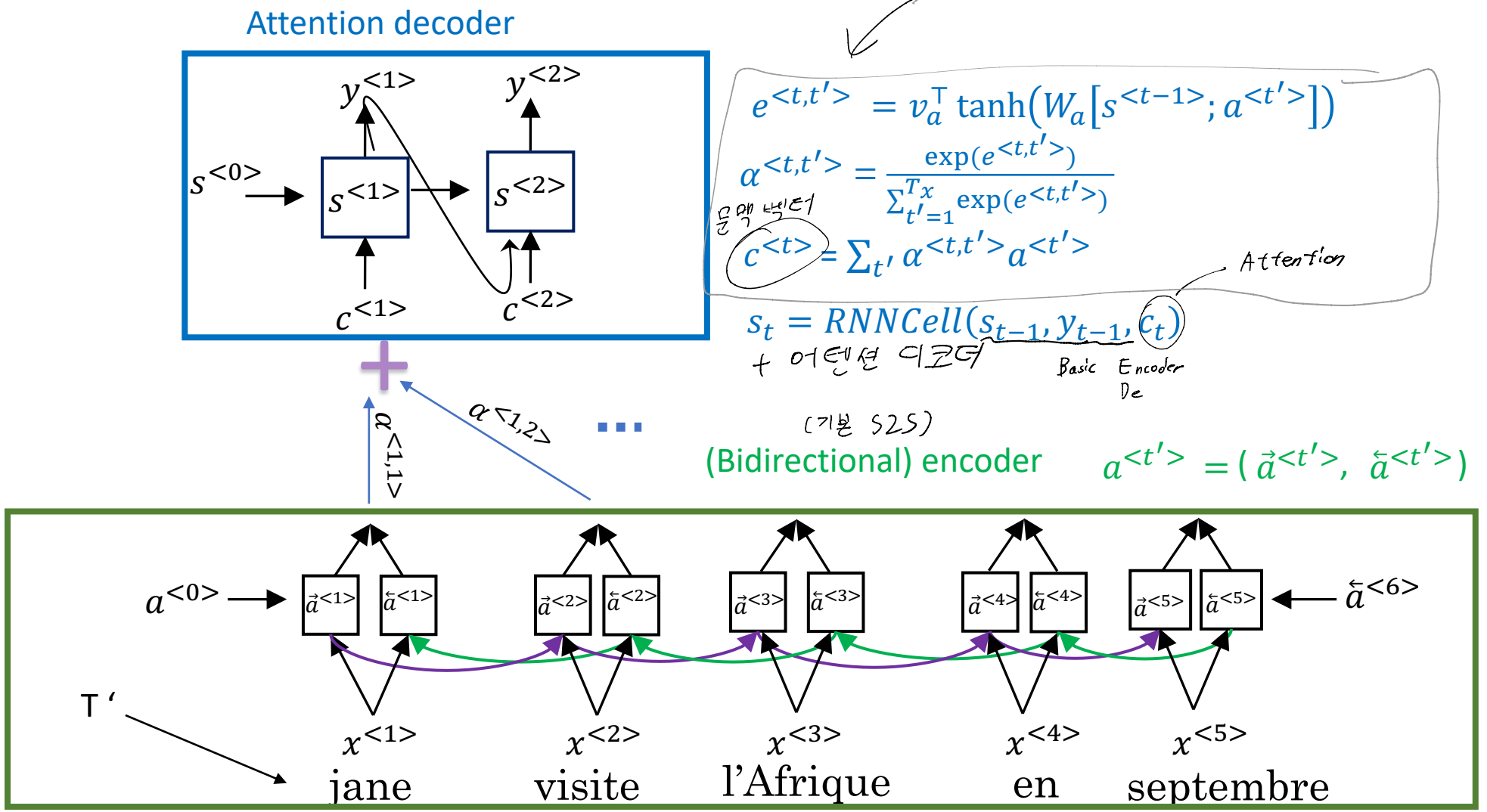
Jane is visiting Africa in September

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$

전처리 단계

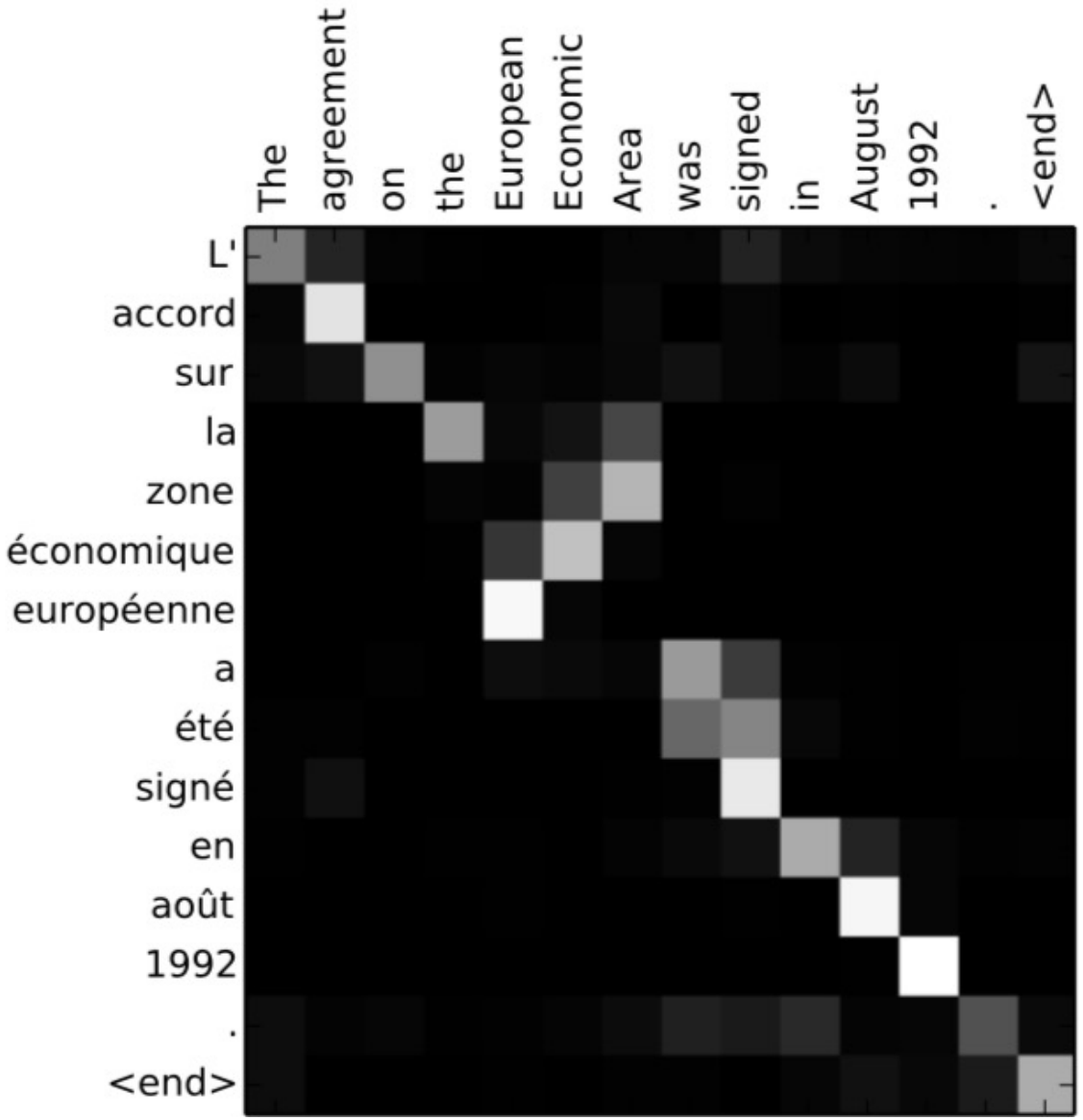


Attention Model



[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

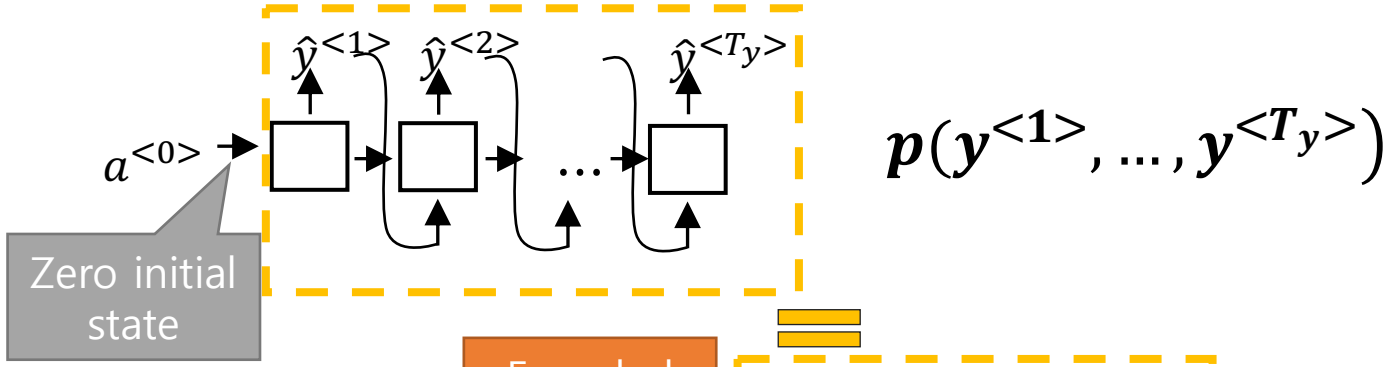
Attention


The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>

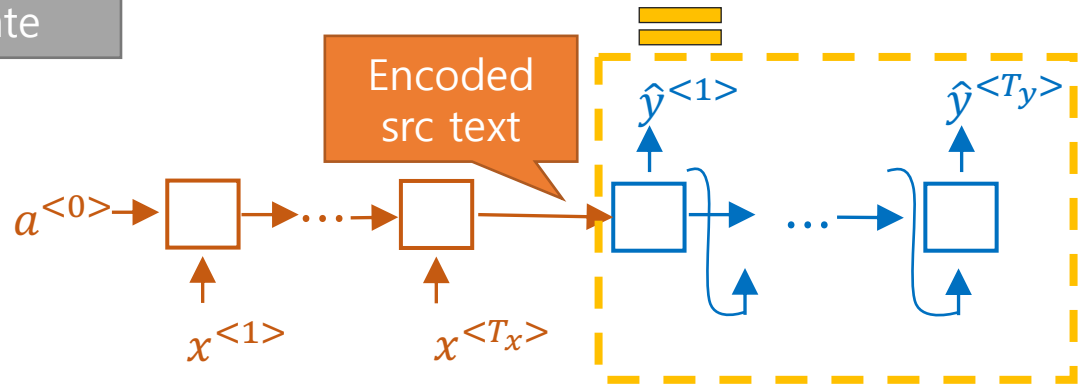
Generating the most likely sentence

Machine Translation as Building a Conditional Language Model

Language model:



Machine translation:



Conditional language model $p(y^{<1>}, \ldots, y^{<T_y>} \mid x^{<1>}, \ldots, x^{<T_x>})$

Finding the Most likely Translation

Jane visite l'Afrique en septembre.

$P(y^{<1>}, \ldots, y^{<T_y>} | x)$

English

French

→ Jane is visiting Africa in September.

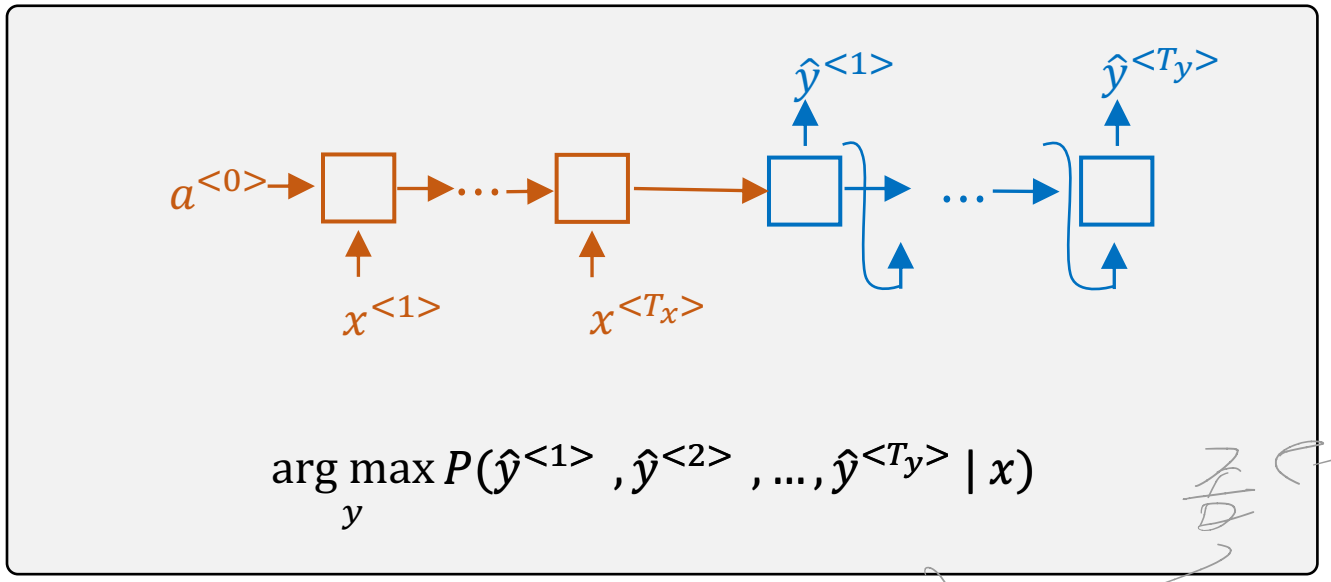
→ Jane is going to be visiting Africa in September.

→ In September, Jane will visit Africa.

→ Her African friend welcomed Jane in September.

$$\arg\max_{y^{<1>},\ldots,y^{<T_y>}} P(\hat{y}^{<1>}, \hat{y}^{<2>}, \ldots, y^{<T_y>} | x)$$

Why not a Greedy Search? 그리디는 안됨



좋은 단어를 필요

Jane is visiting Africa in September.

Jane is going to be visiting Africa in September.

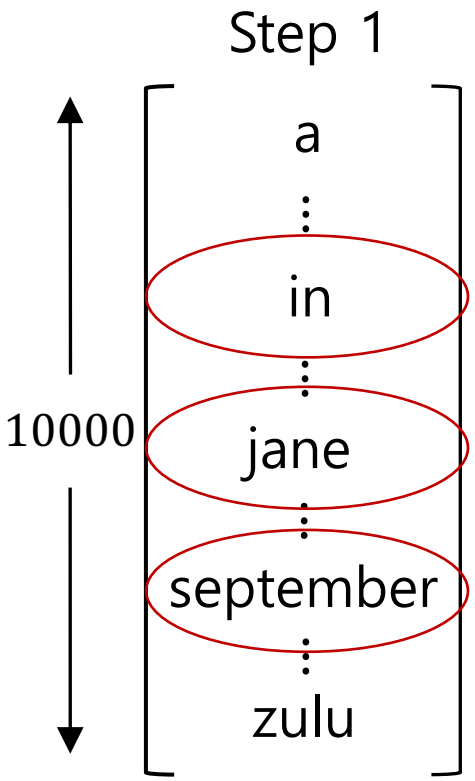
$P(\text{Jane is going} \mid x) > P(\text{Jane is visiting} \mid x)$

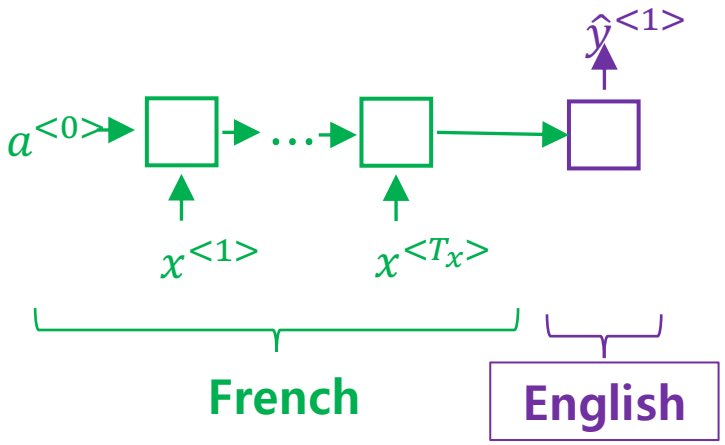
그리드는 이거 골라버림

Beam Search

Beam Search

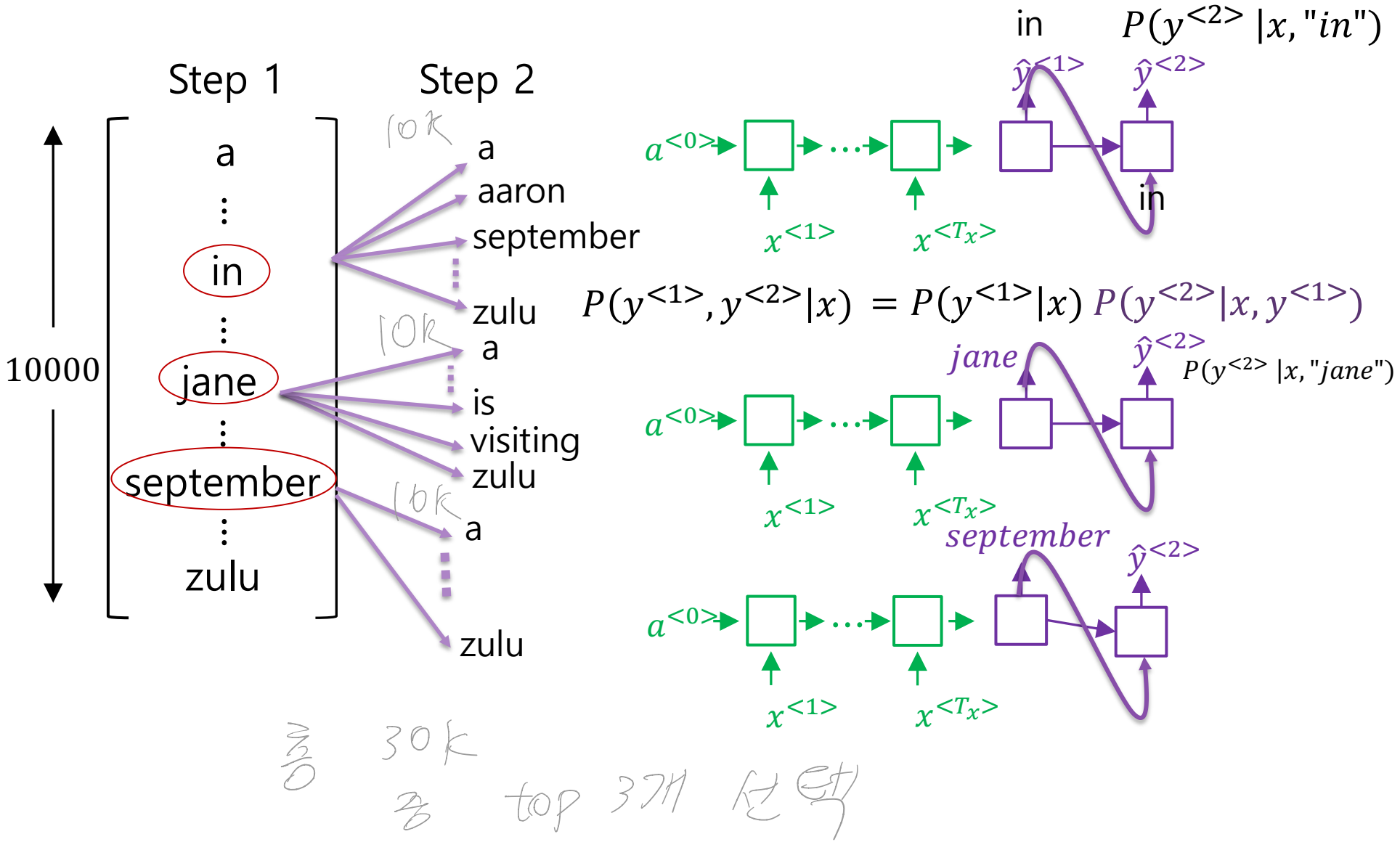
B = 3 (Beam width)



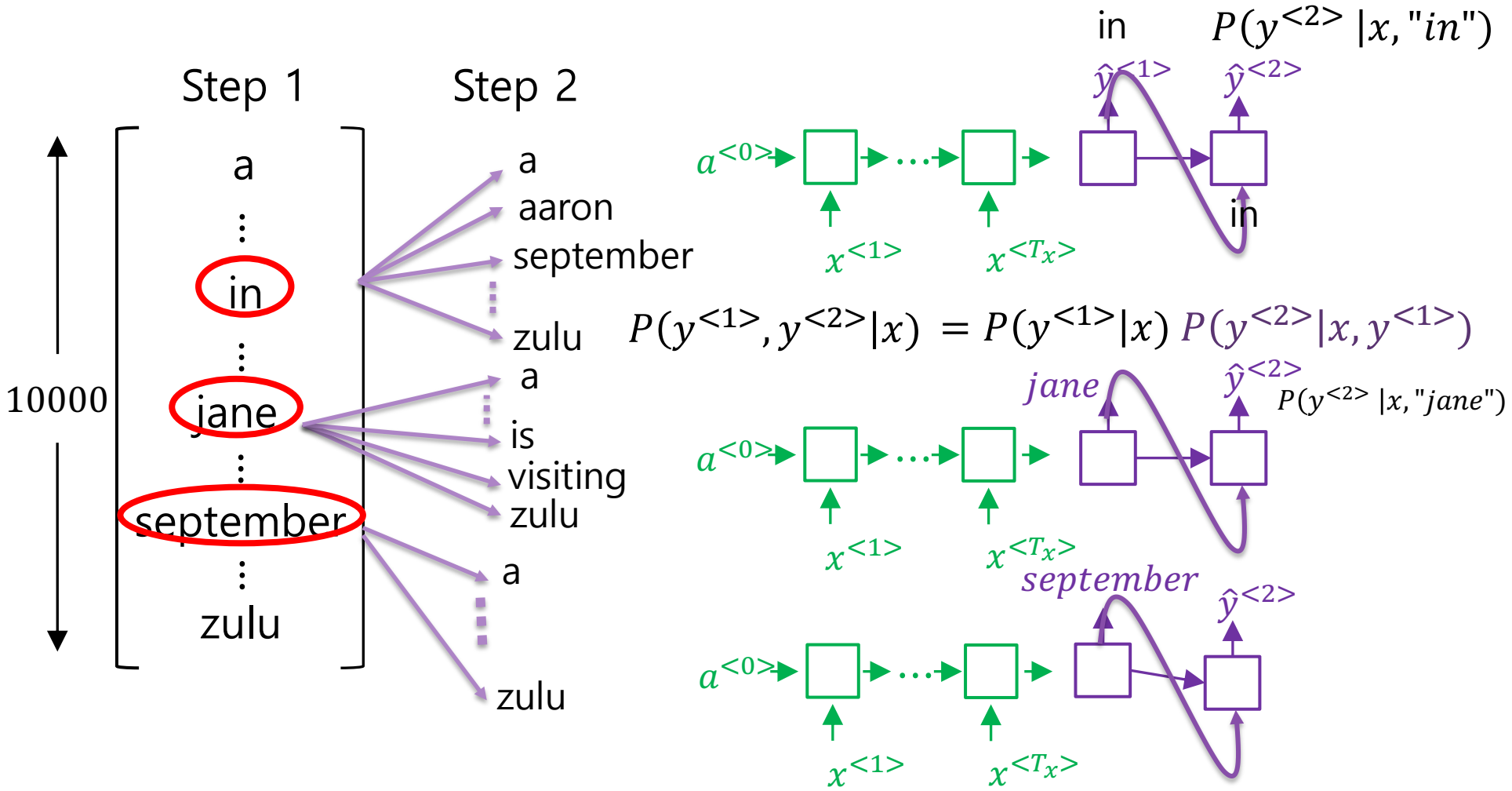


$P(y^{<1>} \mid x)$

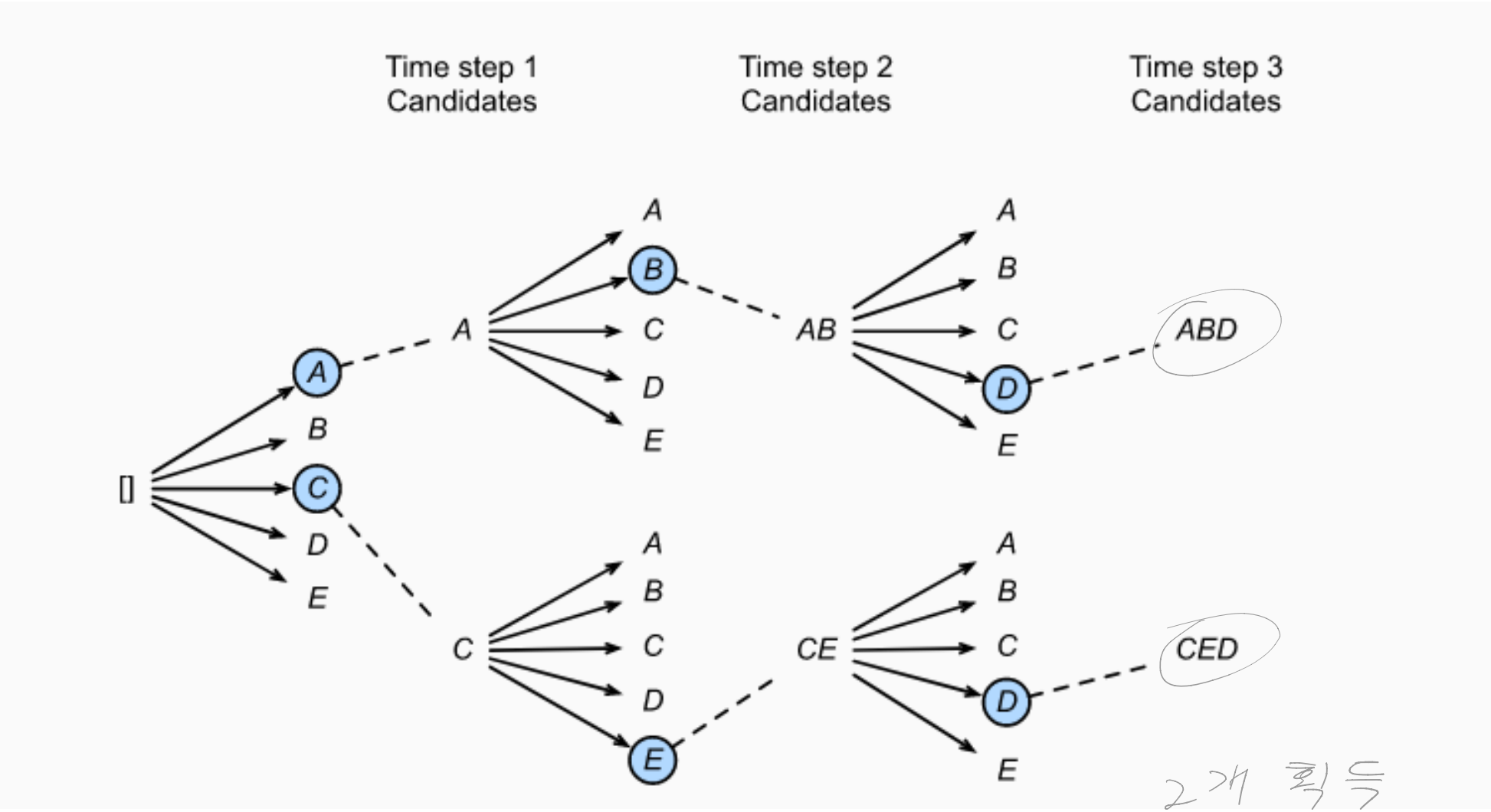
Beam Search (B=3)



Beam Search (B=3)



Beam Search (B=2)



* https://d2l.ai/chapter_recurrent-modern/beam-search.html

Refinements to Beam Search : Length Normalization

길이 정규화

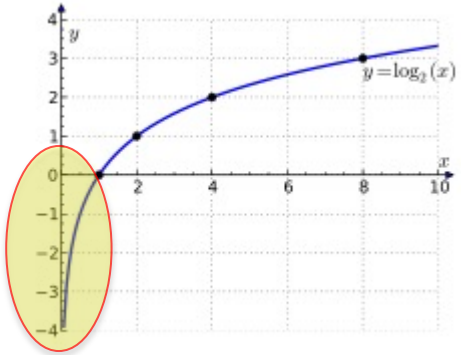
$$P(y^{<1>}, \dots, y^{<T_y>} | x) = P(y^{<1>} | x)\, P(y^{<2>} | x, y^{<1>}) \dots P(y^{<T_y>} | x, y^{<1>}, \dots, y^{<T_y-1>})$$

$$\arg\max_y \prod_{t=1}^{T_y} P(y^{<t>} | x, y^{<1>}, \dots, y^{<t-1>})$$

log P

$$\arg\max_y \sum_{t=1}^{T_y} \log P(y^{<t>} | x, y^{<1>}, \dots, y^{<t-1>})$$

<1



log 값 0보다 작음 → 길이가 짧은게 골라질 확률 높음

$$\frac{1}{T_y^{\alpha}} \sum_{t=1}^{T_y} \log P(y^{<t>} | x, y^{<1>}, \dots, y^{<t-1>})$$

정규화

$\alpha = 0.7$ 적절.

$\alpha = 0$? $\alpha = 1$?

short 문장 long 문장

Beam Width B

- The greater the beam width, the fewer states are pruned
  - Large B: Better result, slower
  - Small B: Worse result, faster

$B = \infty$ → 퀄리티 ↑

$B = 1$ → 그리디, 시간 효율적

Preview - Transformers

Writing a deep learning paper

- You create a new deep learning model and write a paper about it. What content should it include?

- 모델 구조

- 경험

- 문제 정의 (x → y)

Transformers: The Architecture Behind ChatGPT

ChatGPT

- Task
  - Chatbot
- The architecture
  - Transformer (we'll study this next week)
- Training
  - Unsupervised Pretraining. (our last topic)
  - Reinforcement Learning