MxNet Gluon

Basics, Computer Vision, NLP (and even more NLP)

Part VI (Beam Search)

Leonard Lausen
Haibin Lin
Alex Smola

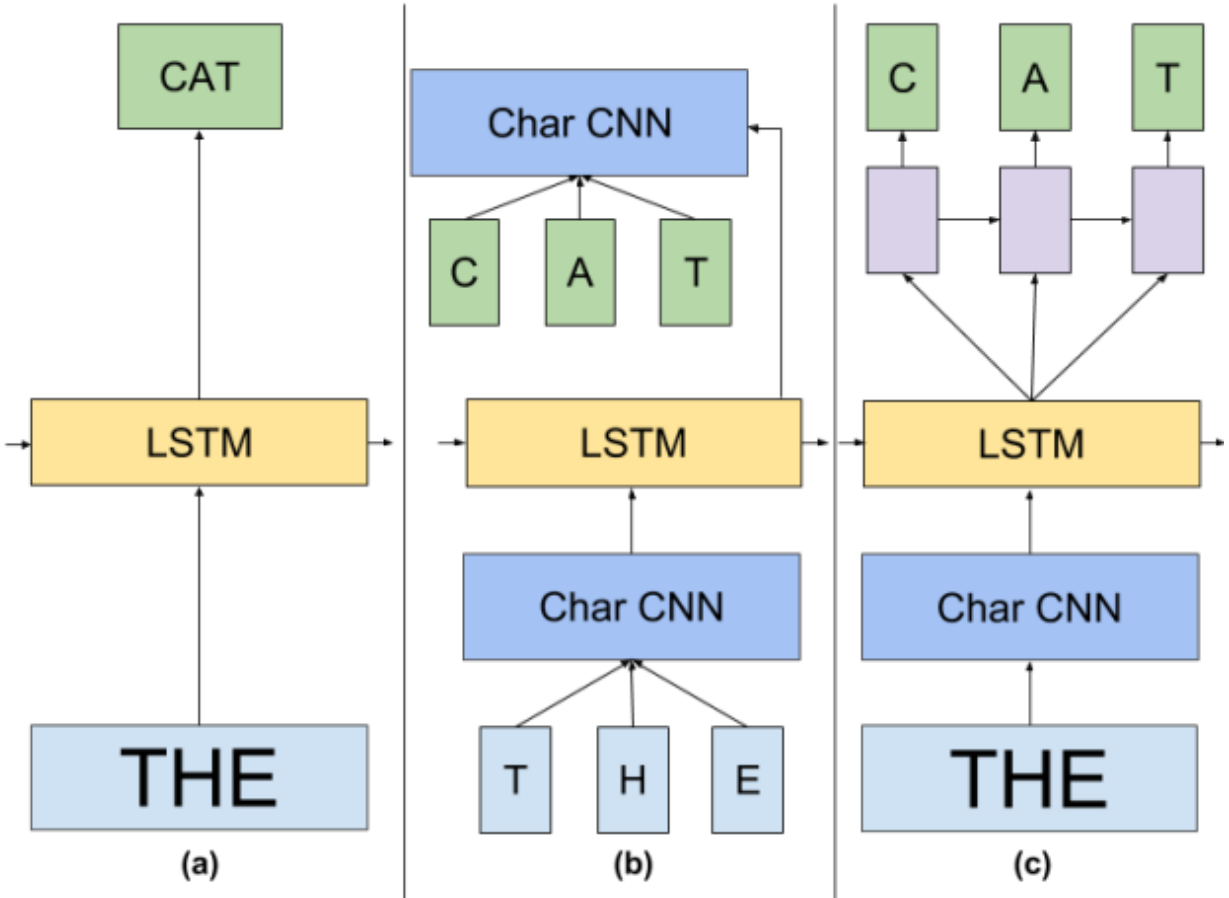
aws

Outline

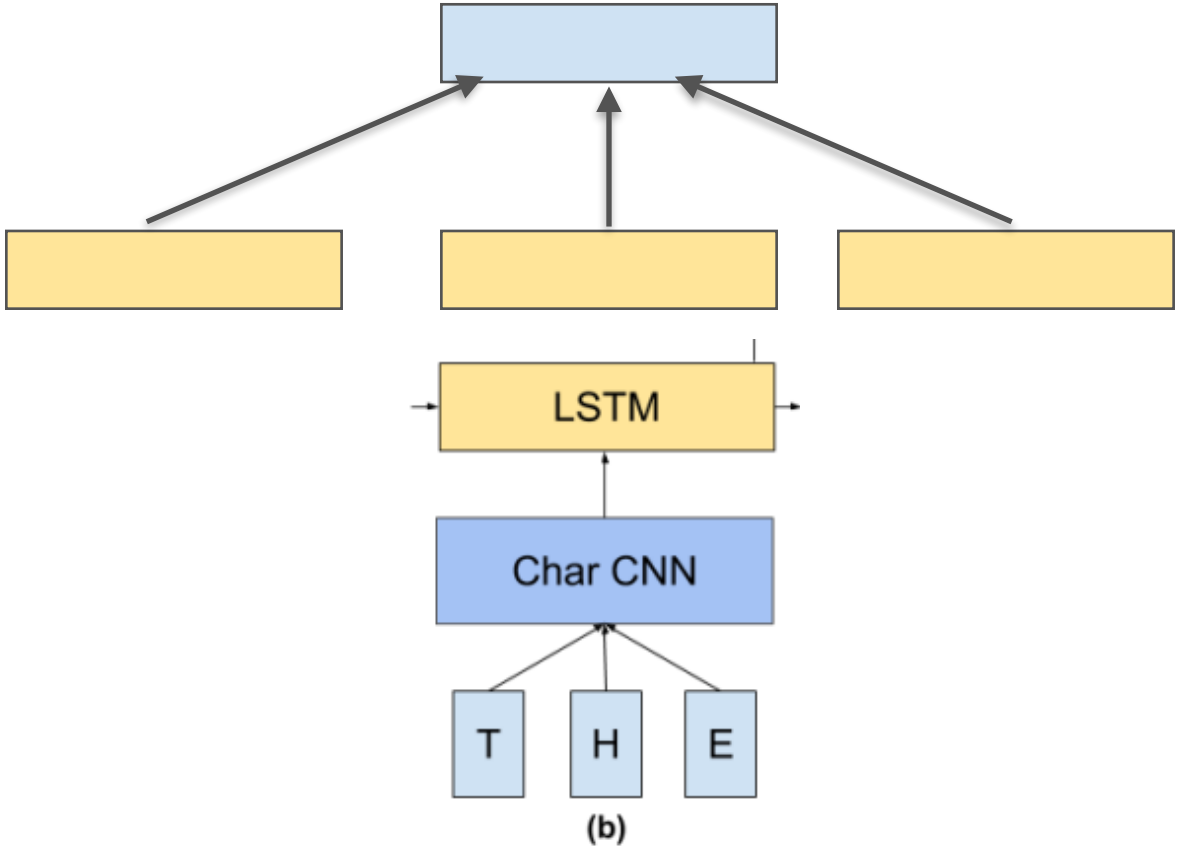
8:30-9:15	Installation and Basics (NDArray, AutoGrad, Libraries)
9:15-9:30	Neural Networks 101 (MLP, ConvNet, LSTM, Loss, SGD) - Part I
9:30-10:00	Break
10:00-10:30	Neural Networks 101 (MLP, ConvNet, LSTM, Loss, SGD) - Part II
10:30-11:00	Computer Vision 101 (Gluon CV)
11:00-11:30	Parallel and distributed training
11:30-12:00	Data I/O in NLP (and iterators)
12:00-13:30	Break
13:30-14:15	Embeddings
14:15-15:00	Language models (LM)
15:00-15:30	Sequence Generation from LM
15:30-16:00	Break
16:00-16:15	Sentiment analysis
16:15-17:00	Transformer Models & machine translation
17:00-17:30	Questions

Language Model but what now?

Repurposing the LM for Sentiment Analysis



Repurposing the LM for Sentiment Analysis



Lots of space for improvement

- Attention model
- Hierarchical attention model
- Bidirectional sequence model (hindsight is 20/20)
  (This is great, he said sarcastically)
- Tune embeddings for sentiment estimation

- Semisupervised learning
- Localized scores

Language Model but what now?

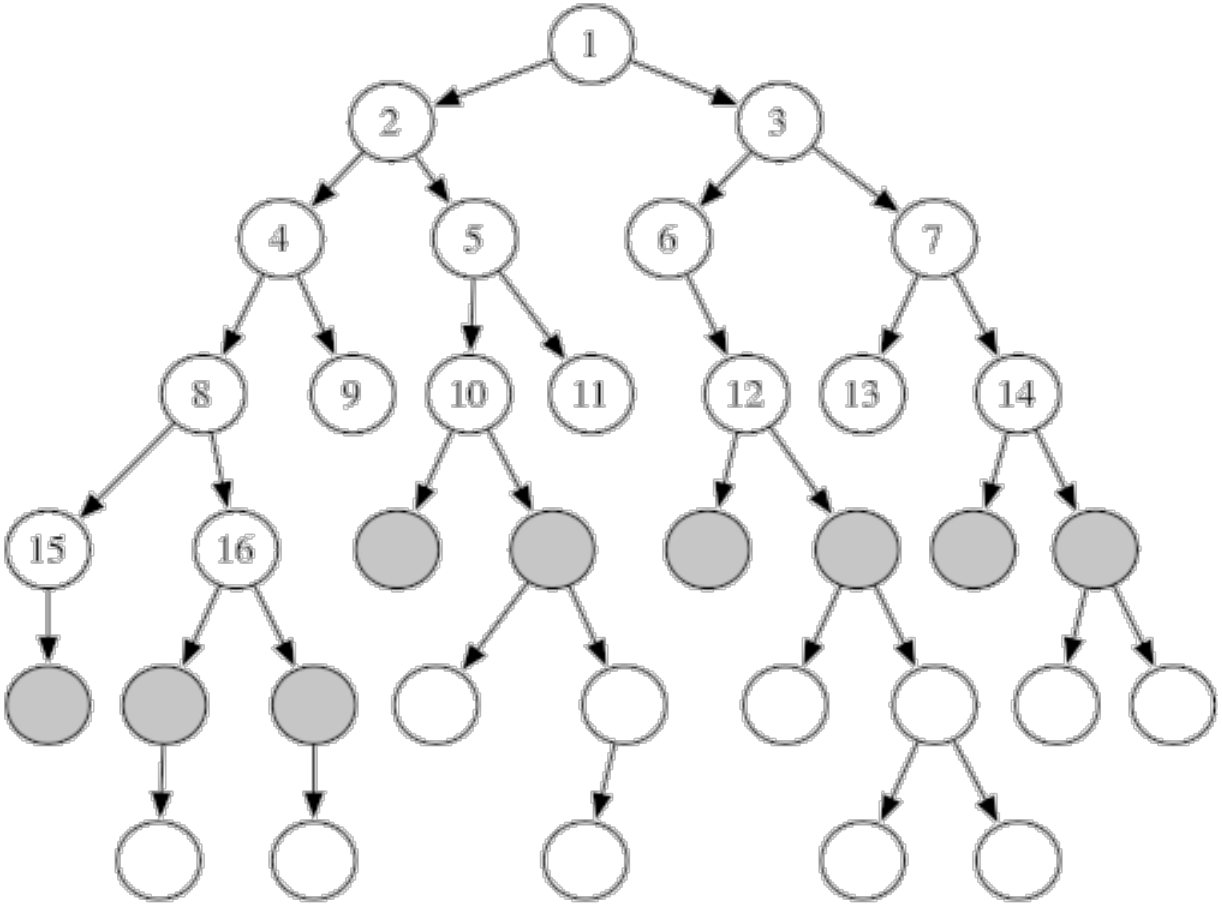
Generating Text

- Language model

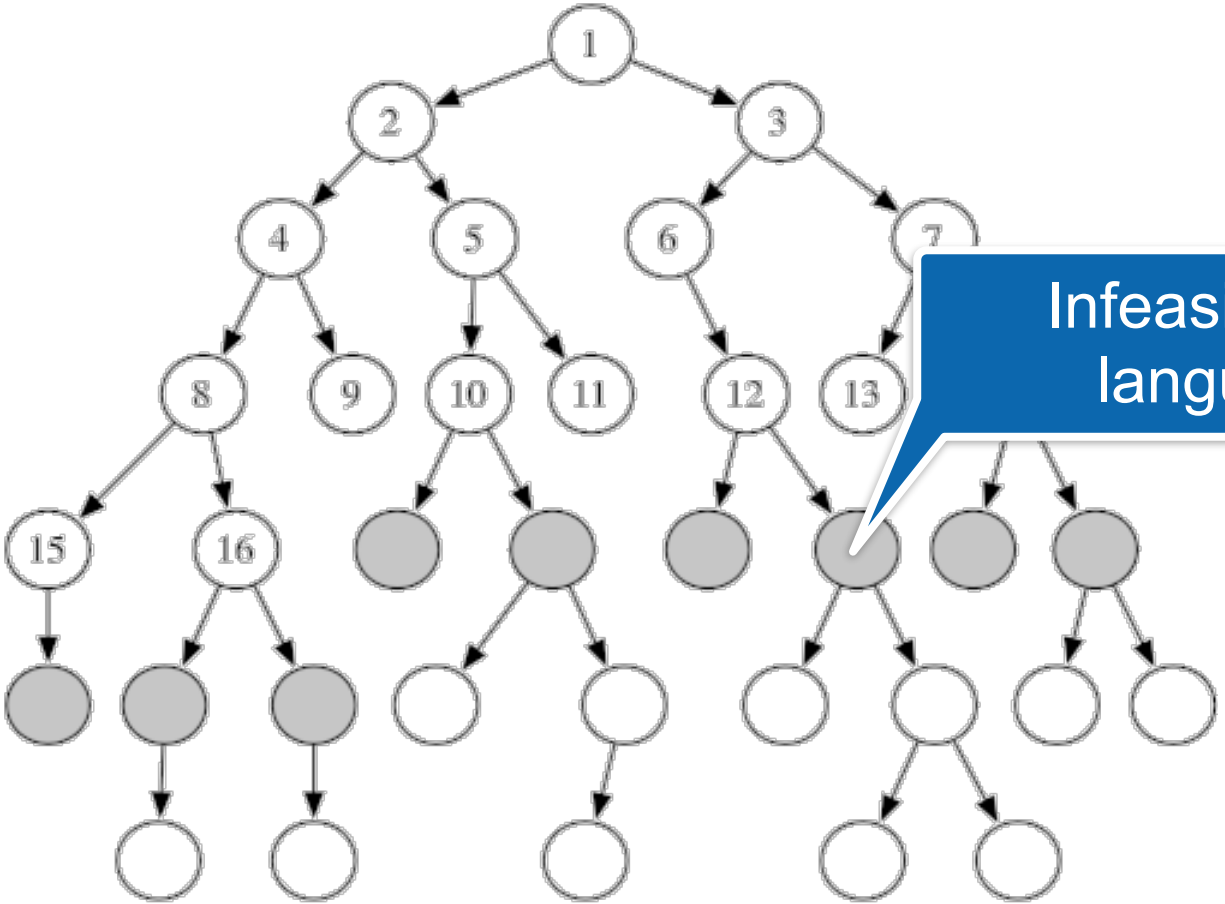
$$p(\text{text}) = \prod_t p(w_t | [w_{t-1} \dots w_1])$$

- Sample from language model … one character at a time
  - Problem - In practice LSTM (or whatever) is not a great approximation, so sampling will not give good text
  - Often want the most likely text (e.g. for translation)
- Need to **search** over lots of possible sequences

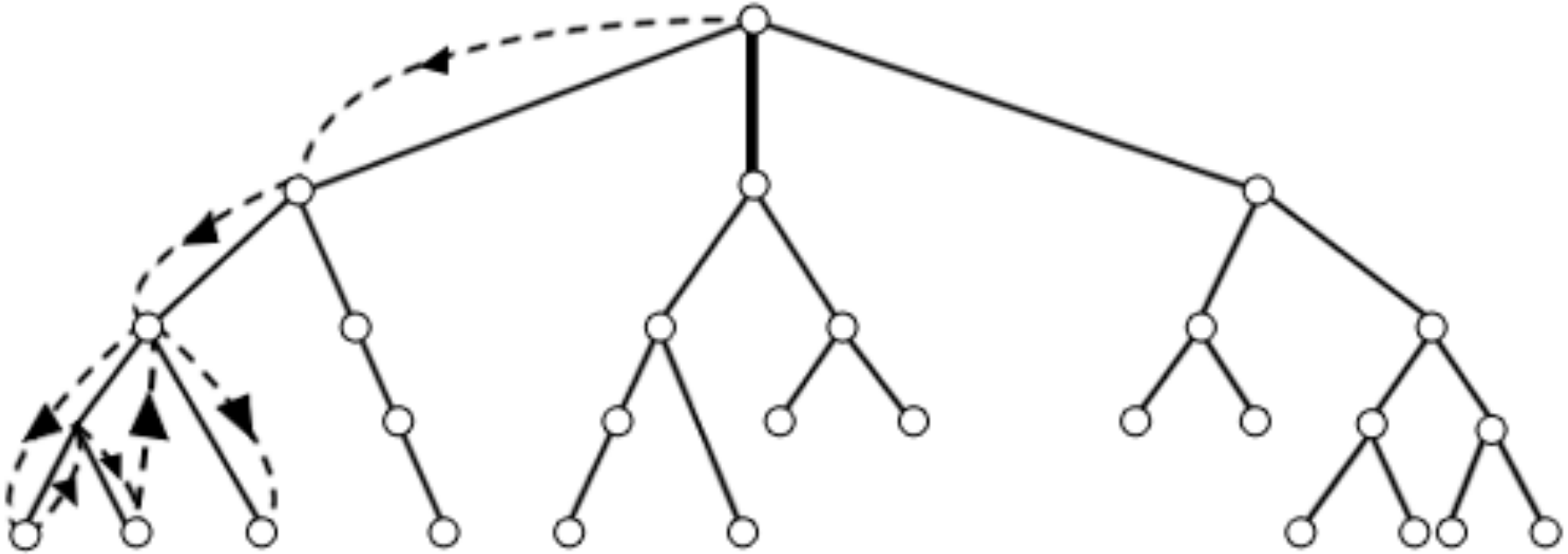
Breadth first search



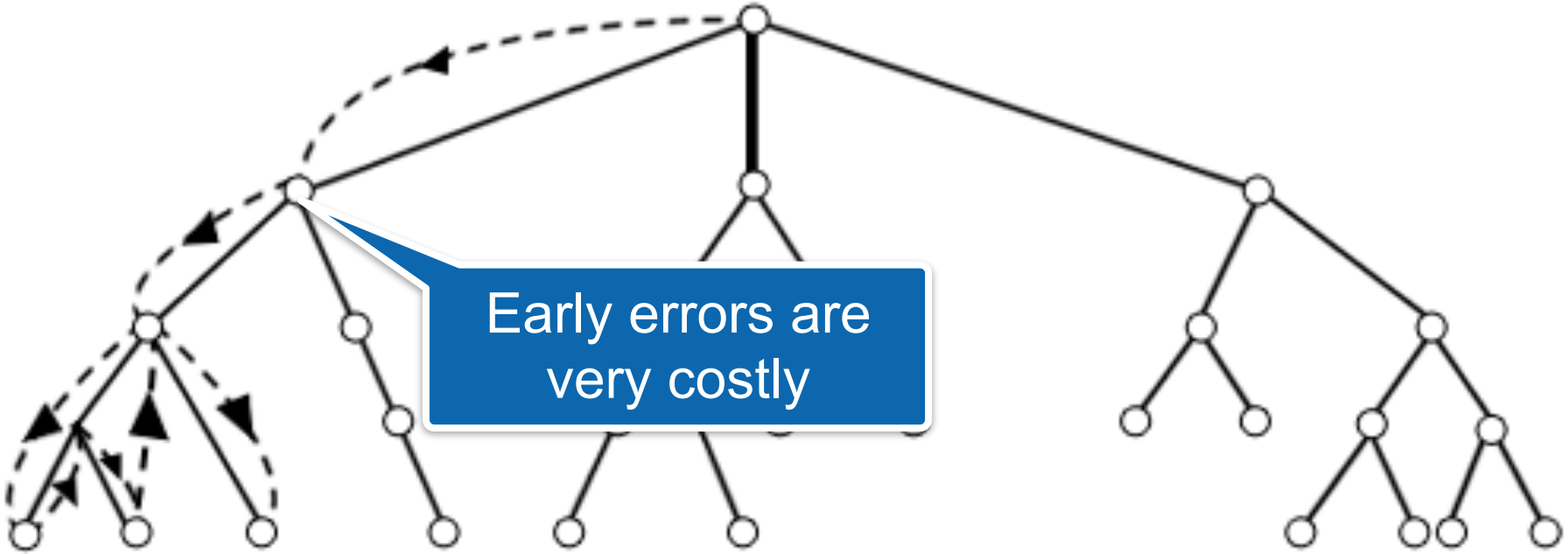
Breadth first search



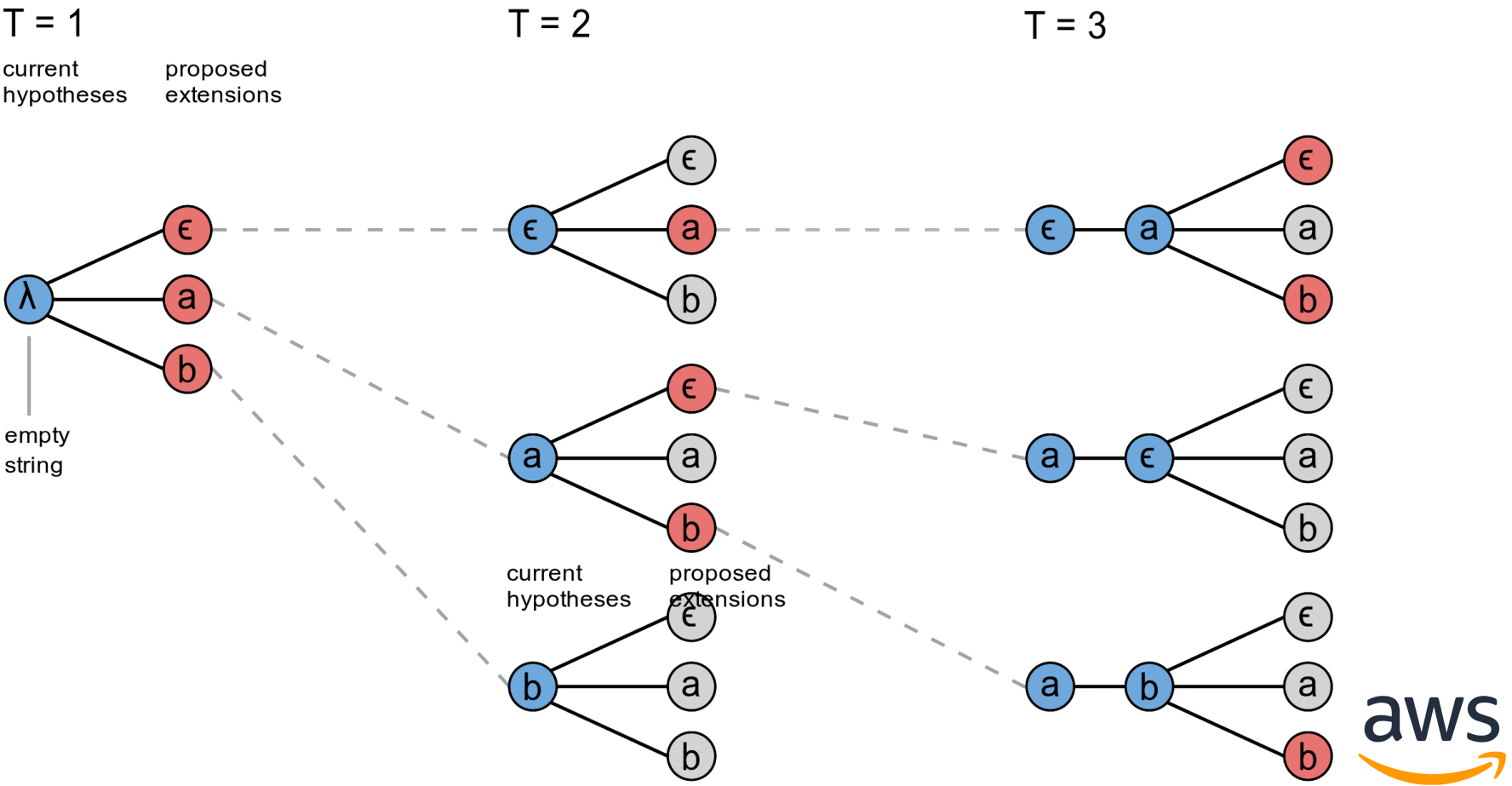
Depth first search



Depth first search



Beam Search


T = 1
current hypotheses
proposed extensions
T = 2
T = 3
λ
ϵ
a
b
empty string
current hypotheses
proposed extensions

Beam Search

- Language model

$$p(\text{text}) = \prod_t p(w_t | [w_{t-1} \dots w_1])$$

- Find top-k paths
  - Start with empty sequence
  - Find top-k extensions of the k sequences (drop rest)
- Beam sampling
  - Similar but sample best extensions

Goldilocks

- **Avoid pathological cases** (Wu et al, 2016)
  - “”
  - “La La La La La La La …”
  - Partial translations in machine translation
- Length penalty, such as $(l+5)^{\alpha}$ to normalize for variable segment lengths
- Submodular Coverage penalty avoids missing segments

$$\sum_i \log \min\left(\sum_j \alpha_{ij}, 1\right)$$