Language Modeling

Instructor: Tom Ko

The basic

- https://web.stanford.edu/~jurafsky/slp3/slides/LM_4.pptx

ARPA file format

- It is a n-gram format in text
- All values are logarithm of base 10
- The first value is the conditional probability
- The last value is the backoff weight

- \data\
  ngram 1=5776
  ngram 2=55566

- \1-grams:
  -4.7039757 <unk> 0
  0 <s> -1.505814
  -4.554822 </s> 0
  -1.7441652 the -1.075229
  -2.4278247 book -1.6445116
  -1.6198214 of -1.2298658
  -1.452003 . -4.009832
  -2.7216232 an -0.6389431
  -3.7242882 account -0.75844955
  …
- \2-grams:
  -0.000042455064 . </s>
  -1.9652246 <s> the
  -0.5145896 of the
  -0.429893 upon the
  -1.6545775 taken the
  -0.3712691 from the
  …

Language modeling with neural network

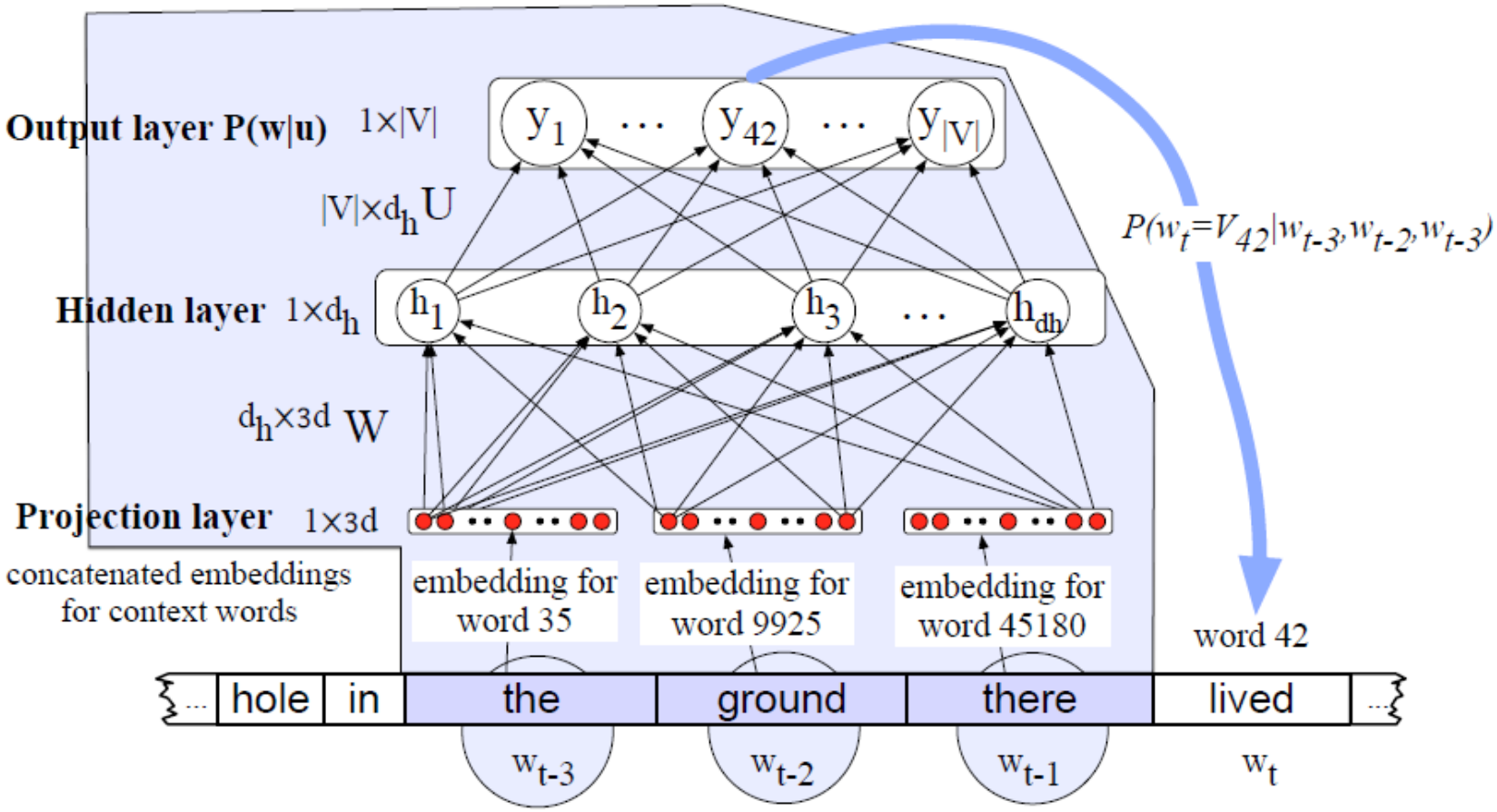
- Have many advantages over the n-gram language models
  - neural language models don't need smoothing
  - can generalize over contexts of similar words
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. Journal of machine learning research, 3(Feb), 1137-1155.

Embeddings

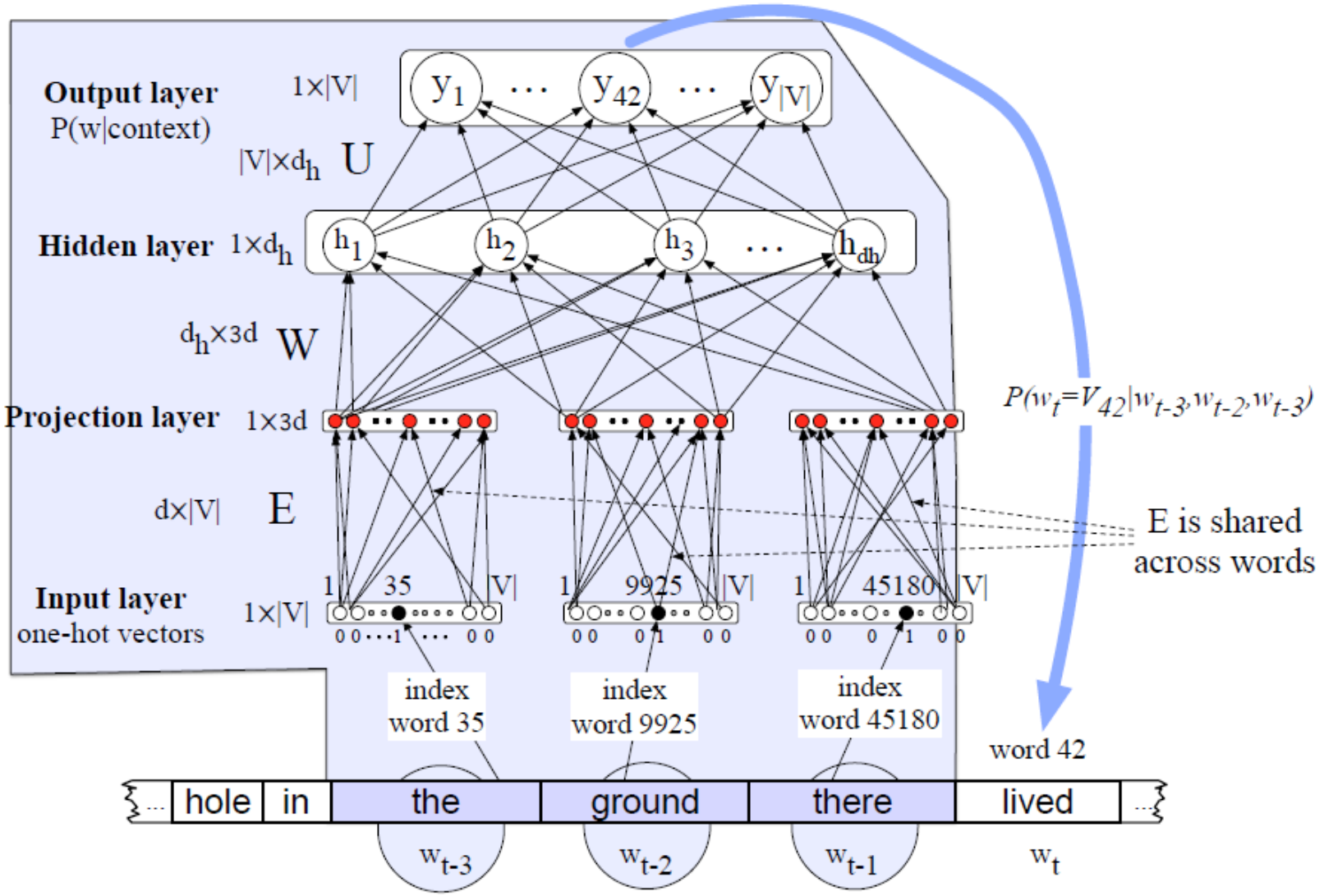
- In neural language models, the prior context is represented by embeddings of the previous words.
- Better in generalizing to unseen data.
- For example, the model sees this senetence in training:
  - I want to buy the house
- In test set, the model encounters a sentence:
  - I wanted to buy the <?>
  - He wanted to sell the <?>

Predicting next word with neural network

- Assume that the embeddings are learnt from other word2vec methods. (pretraining)



Learning the embeddings



Long-distance dependency

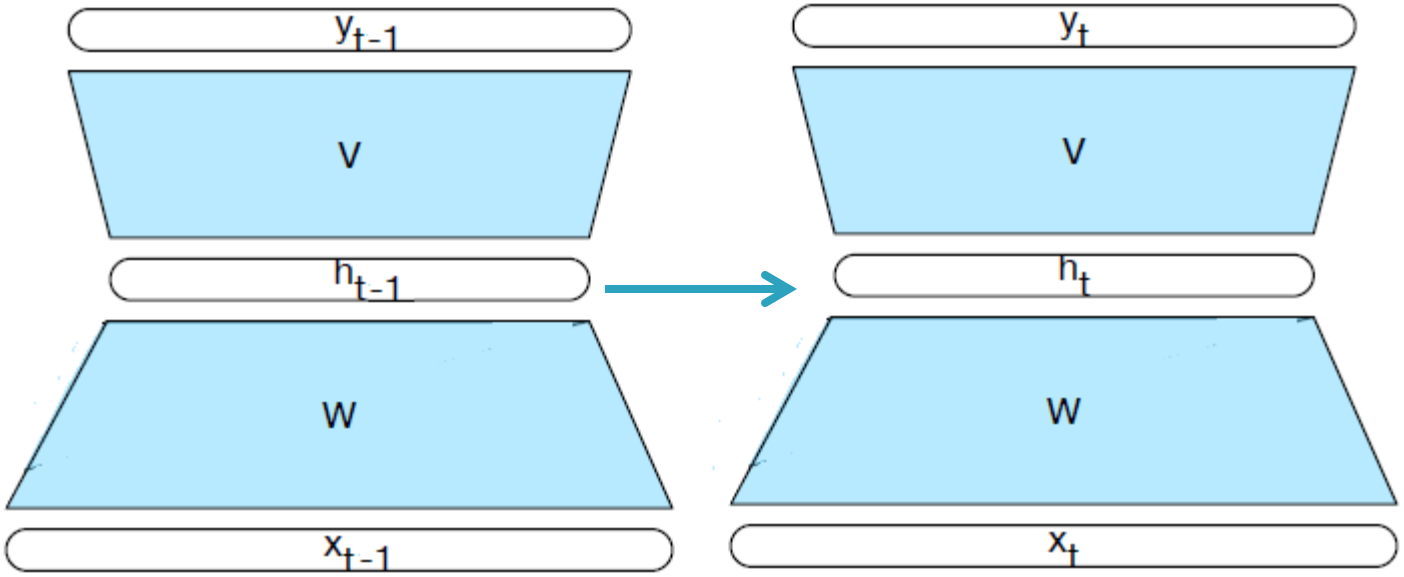
- By using feed forward network, like n-gram, still we have to fix the number of preceding words the model can look at.
- In general this is an insufficient model of language
  - because language has **long-distance dependencies**:

"The computer(s) which I had just put into the machine room on the fifth floor is (are) crashing."

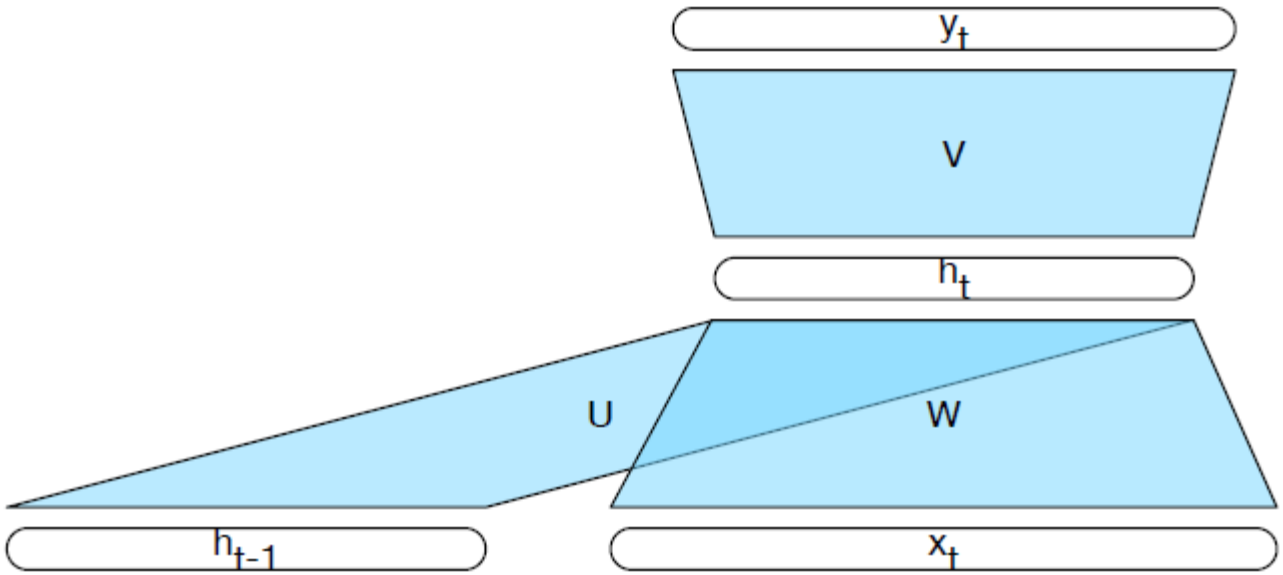
- Solution: Recurrent neural network (RNN)

Recurrent neural network

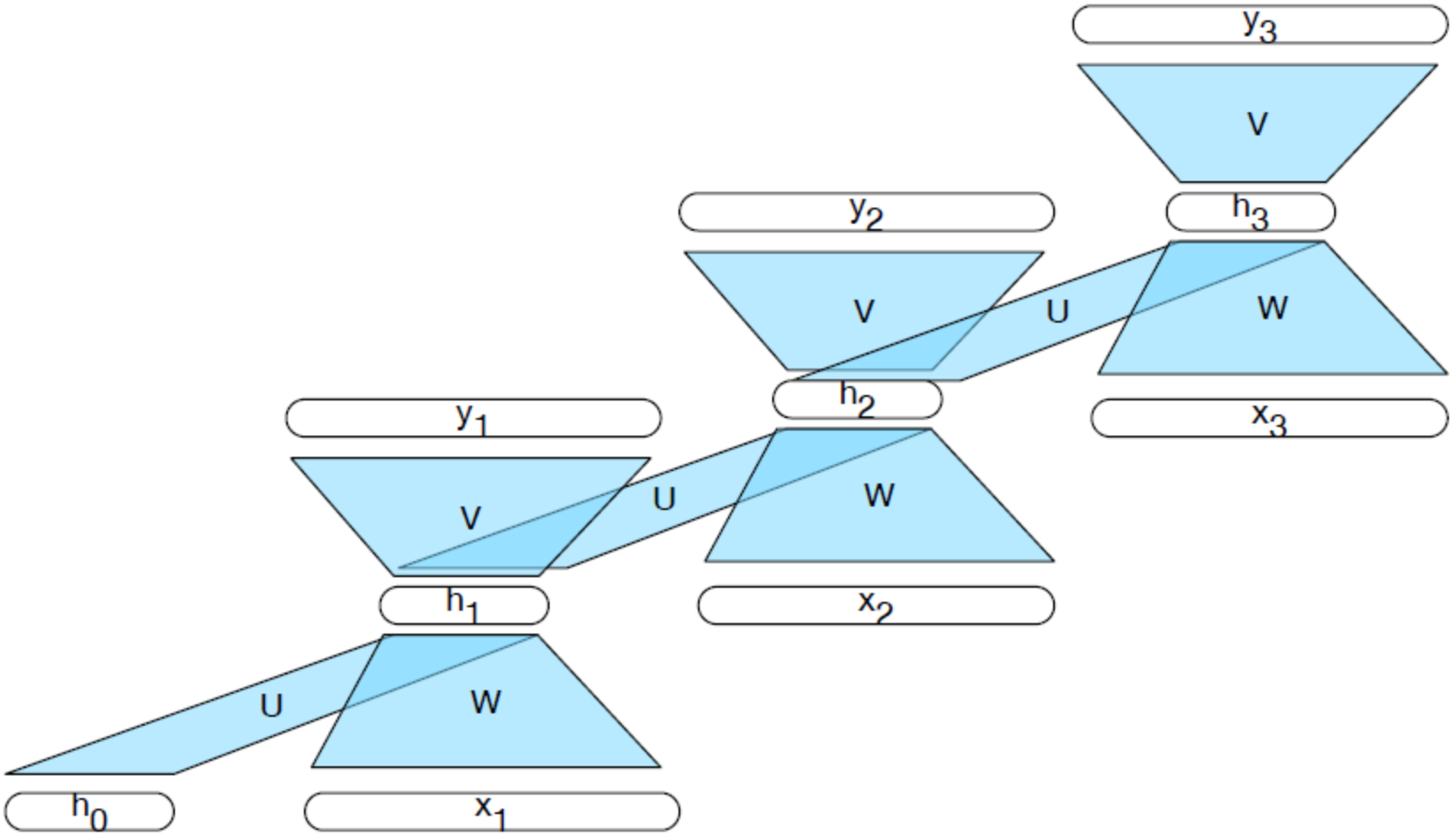
- A recurrent neural network is any network that contains is a cycle within its network connections.



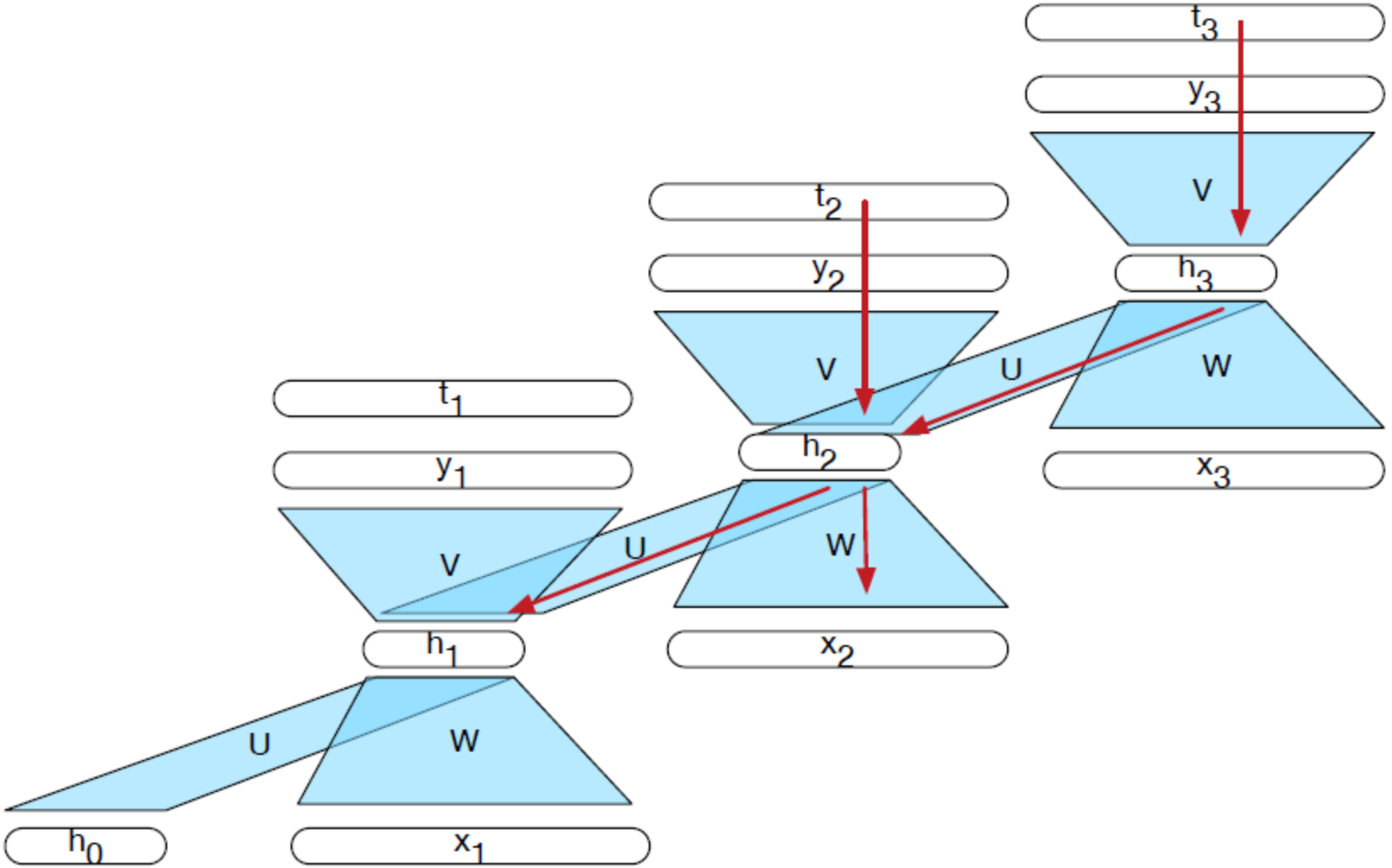
Unrolling the RNN



Unrolling the RNN



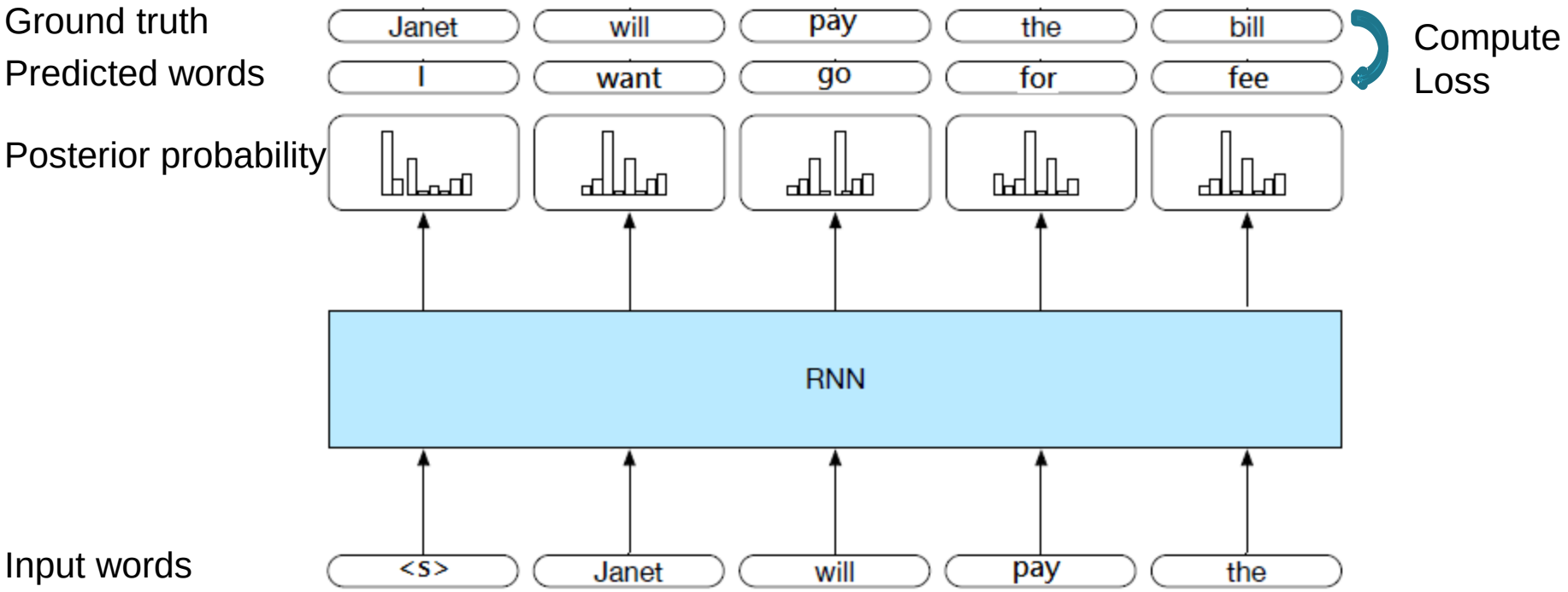
Backpropagation in a RNN



RNN language model

- In feed forward networks, history is represented by context of N−1words, it is limited in the same way as in N-gram backoff models.
- In recurrent networks, history is represented by neurons with recurrent connections – history length is unlimited.
- Also, recurrent networks can learn to compress whole history in low dimensional space, while feed forward networks compress just single word.
- Tomas Mikolov (2010) Recurrent neural network based language model, Interspeech

RNN language model

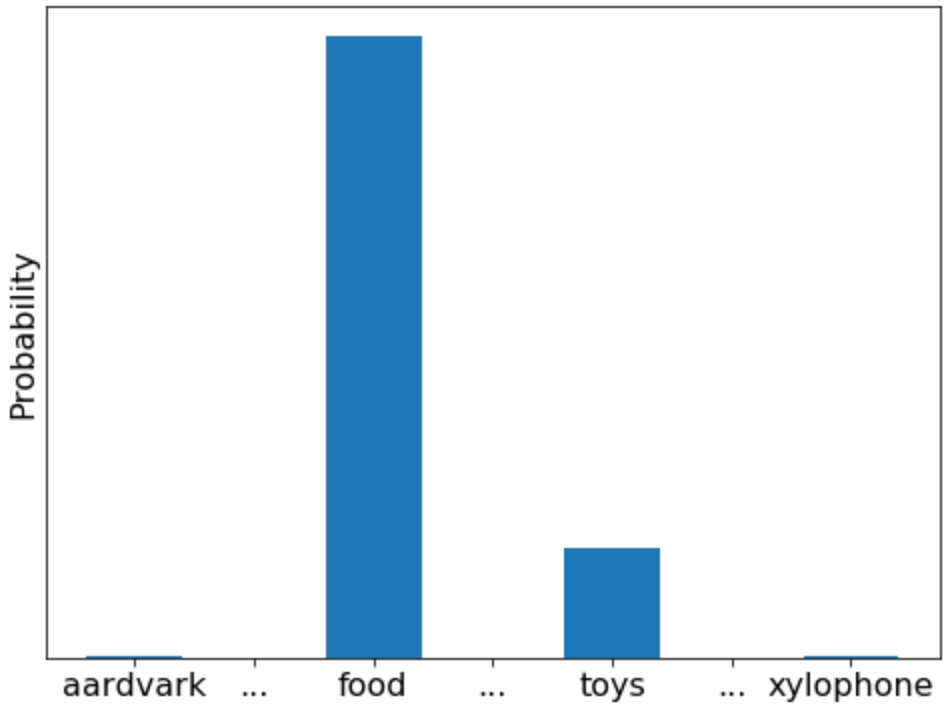


Variations of RNNs

- Long Short Term Memory (LSTM)
- Gated Recurrent Unit (GRU)

Generating new text with LM

- Let's assume we have the sequence [my, cat's, breath, smells, like, cat, ____] and we want to guess the final word. A language model would estimate the probability for every word in the vocabulary:



Generating new text with LM

- *Sampling*:
  - Sample from the conditional word probability distribution. Words that are a better fit are more likely to be selected. For example, we would select the word “food” with probability 62%, “toys” with probability 14%, etc.
- *Greedy*:
  - Always pick the word with the highest probability (aka. argmax). Select “food”.
- *Beam search*:
  - The greedy approach doesn’t always result in the final sequence with the highest overall probability. A beam search keeps track of several probable variants at each step to avoid being led astray by local maxima. Select “food” and “toys”, and reassess what is better when more words have been added.

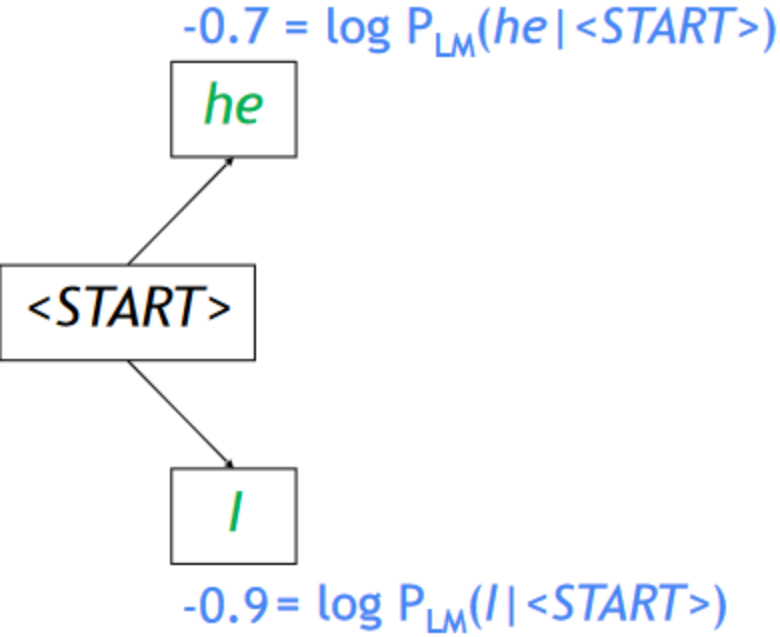
Beam search decoding

- Beam size = 2
- This example is copied from Stanford course CS224n: http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture08-nmt.pdf

<START>

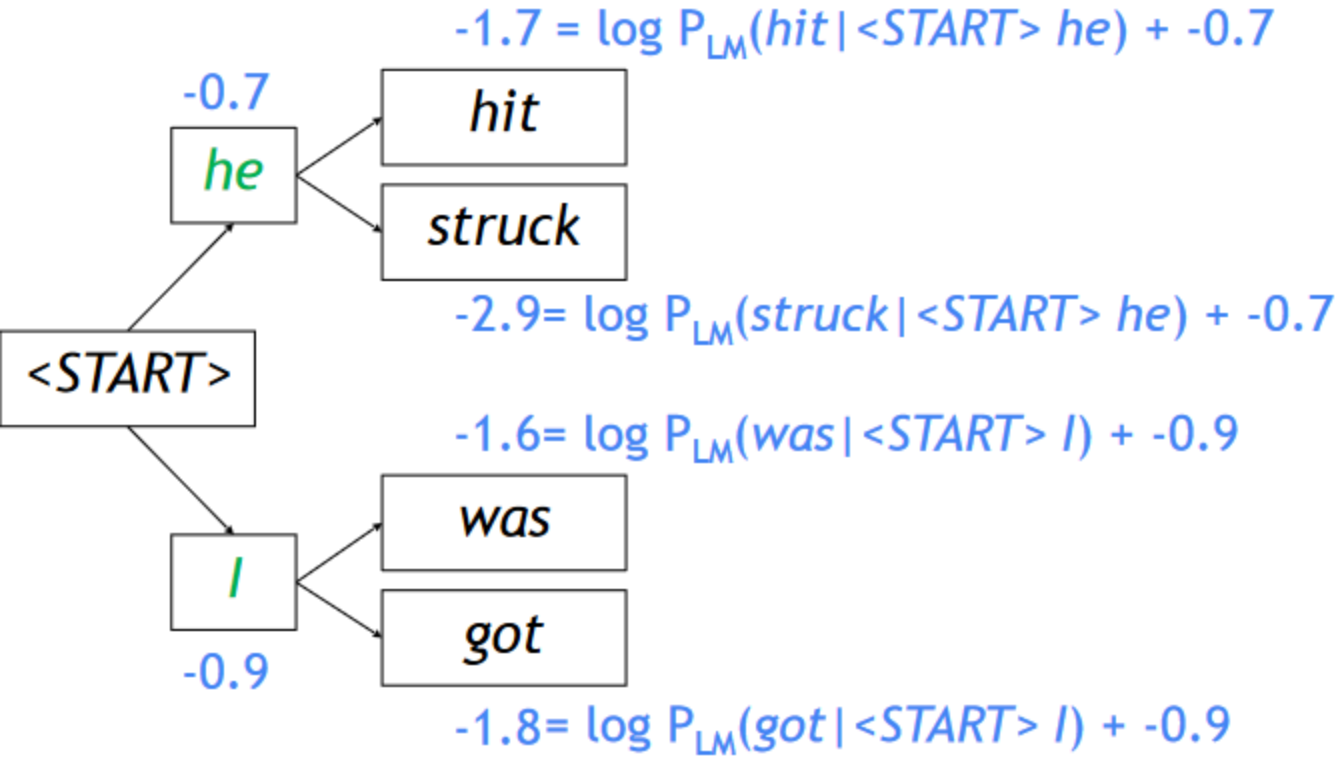
Beam search decoding

- Beam size = 2

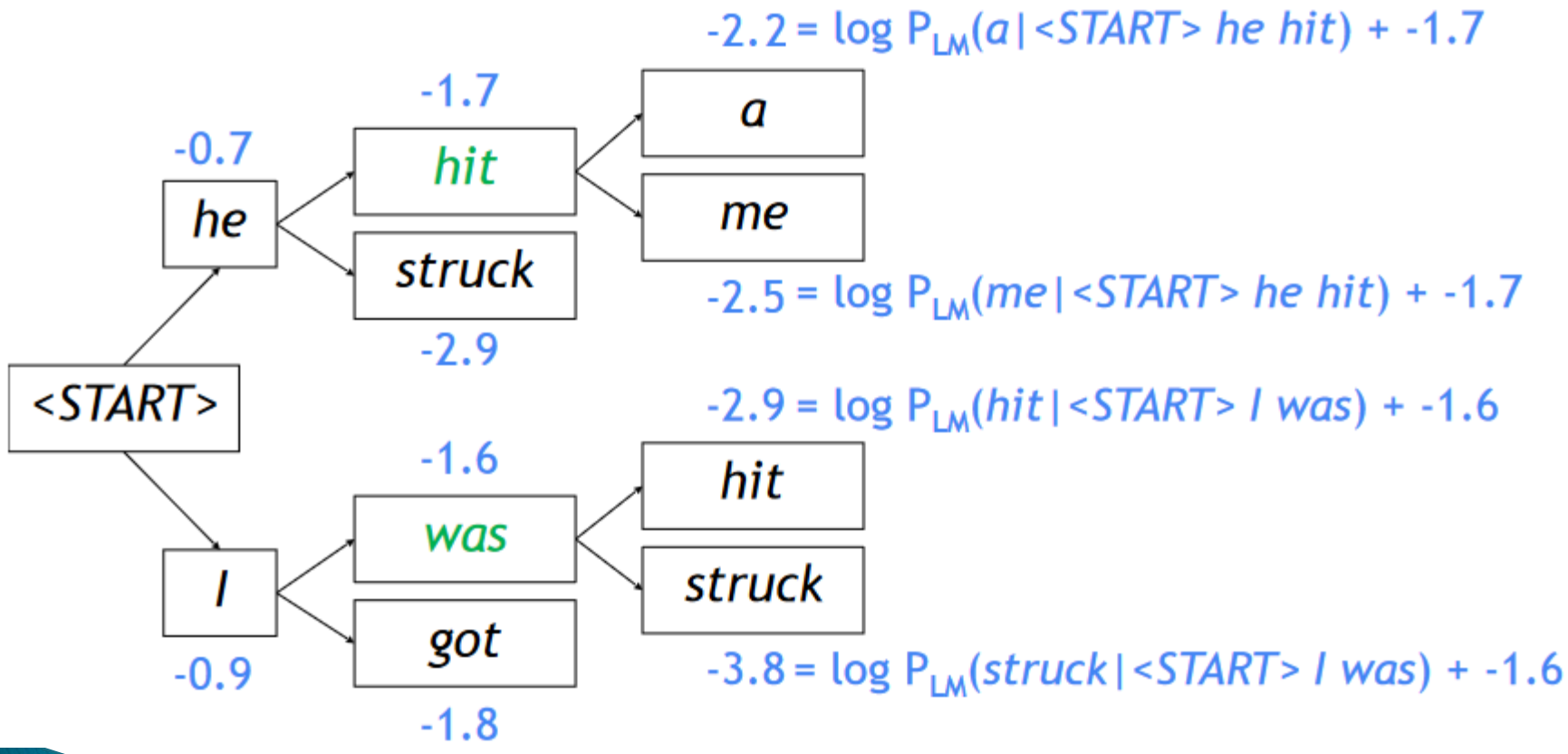


Beam search decoding

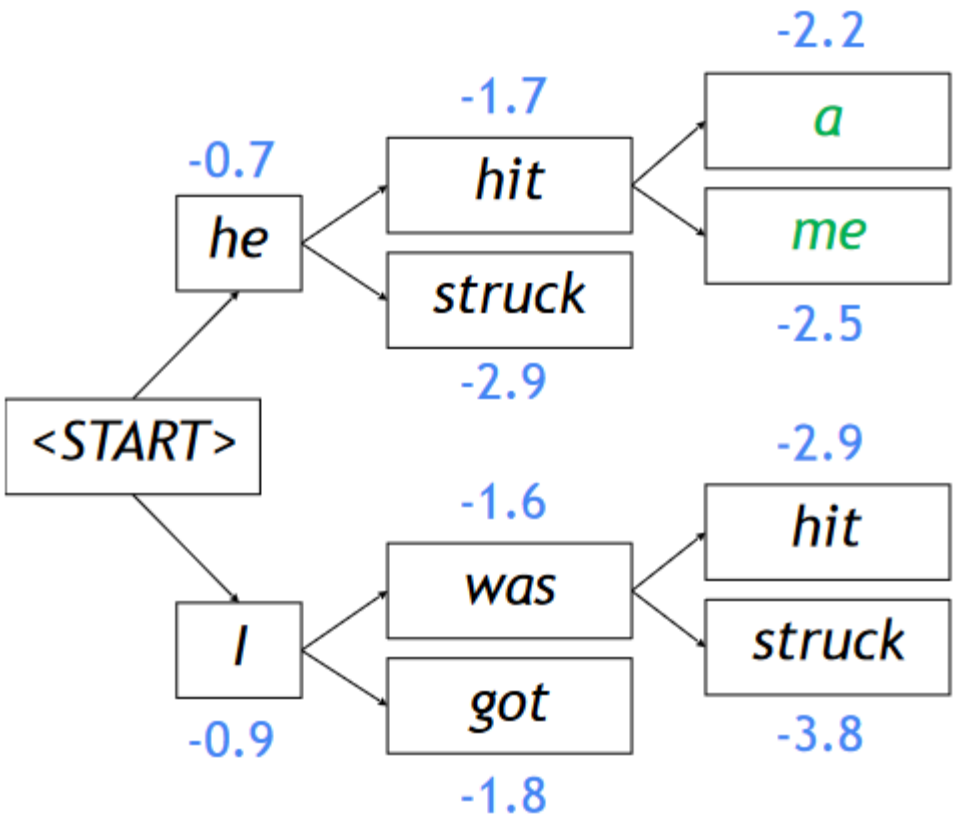
- Beam size = 2



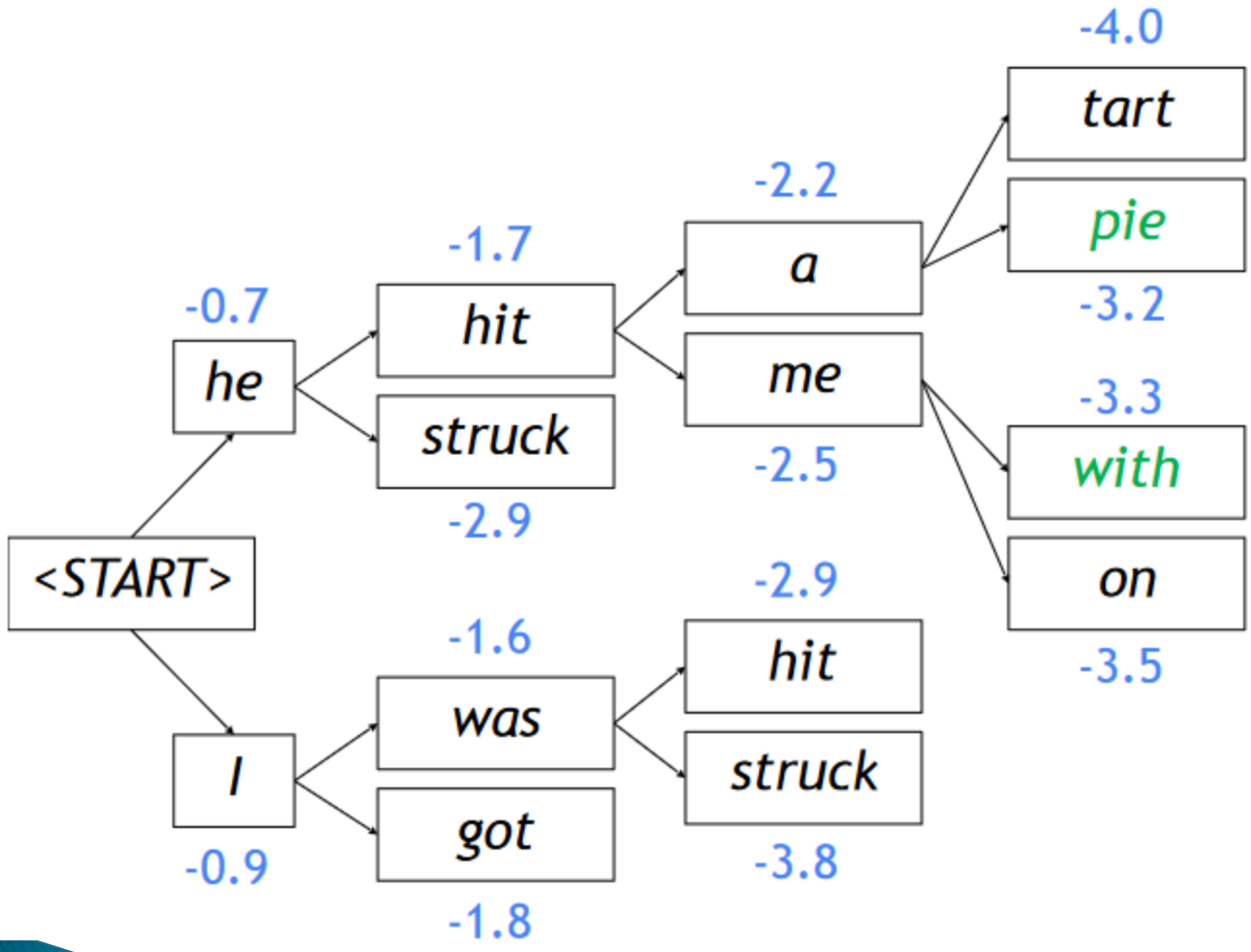
Beam search decoding



Beam search decoding

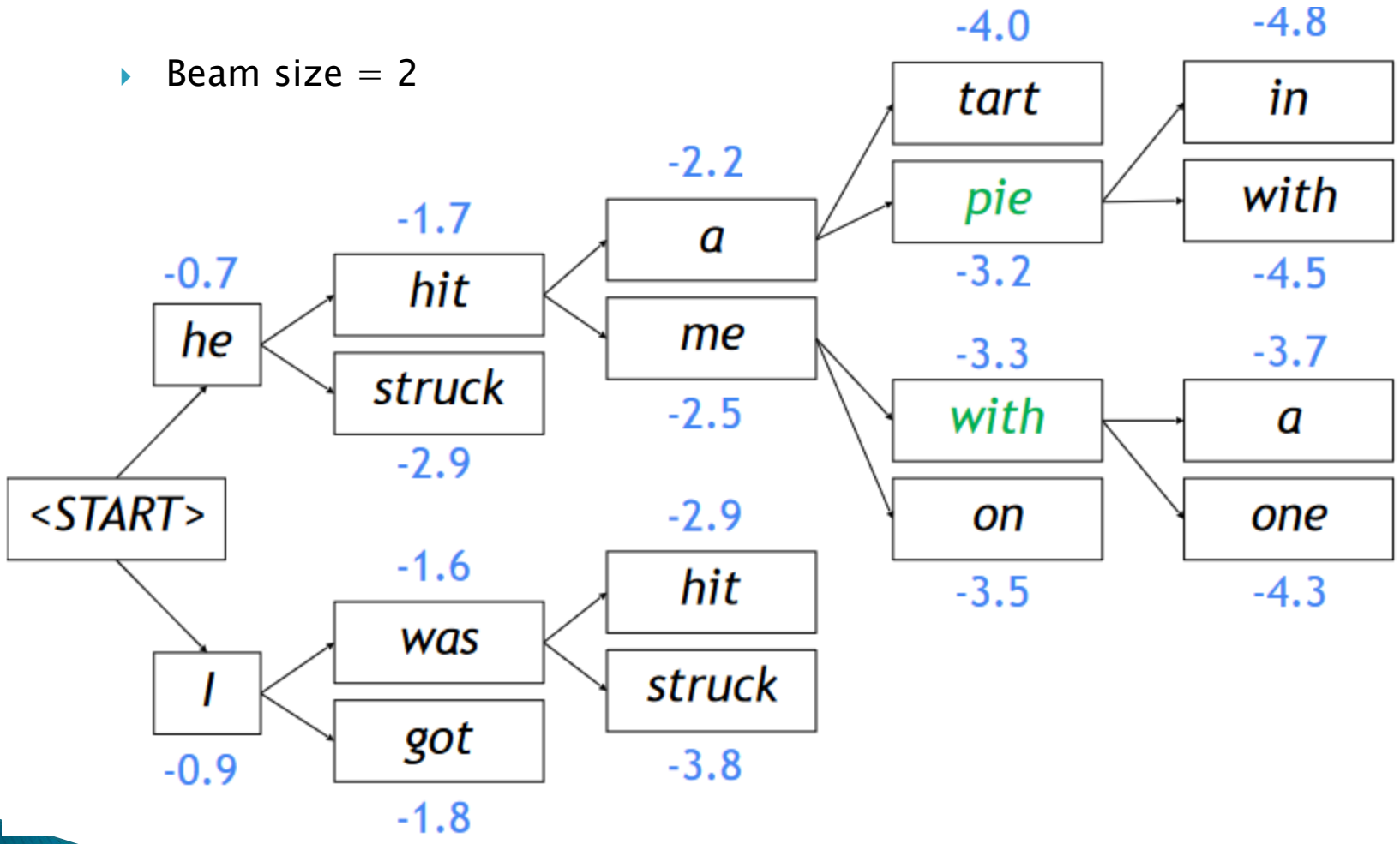


Beam search decoding



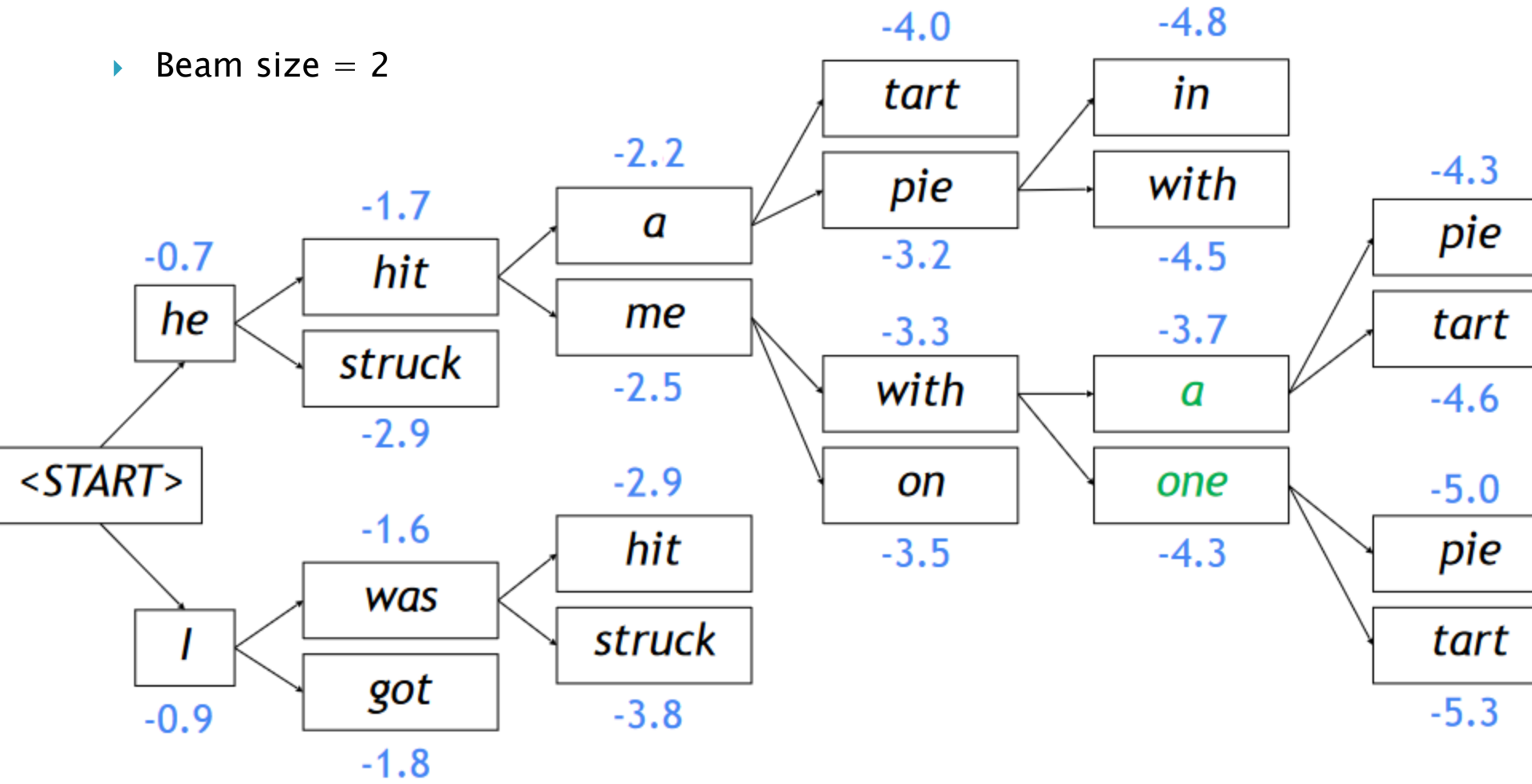
Beam search decoding

- Beam size = 2



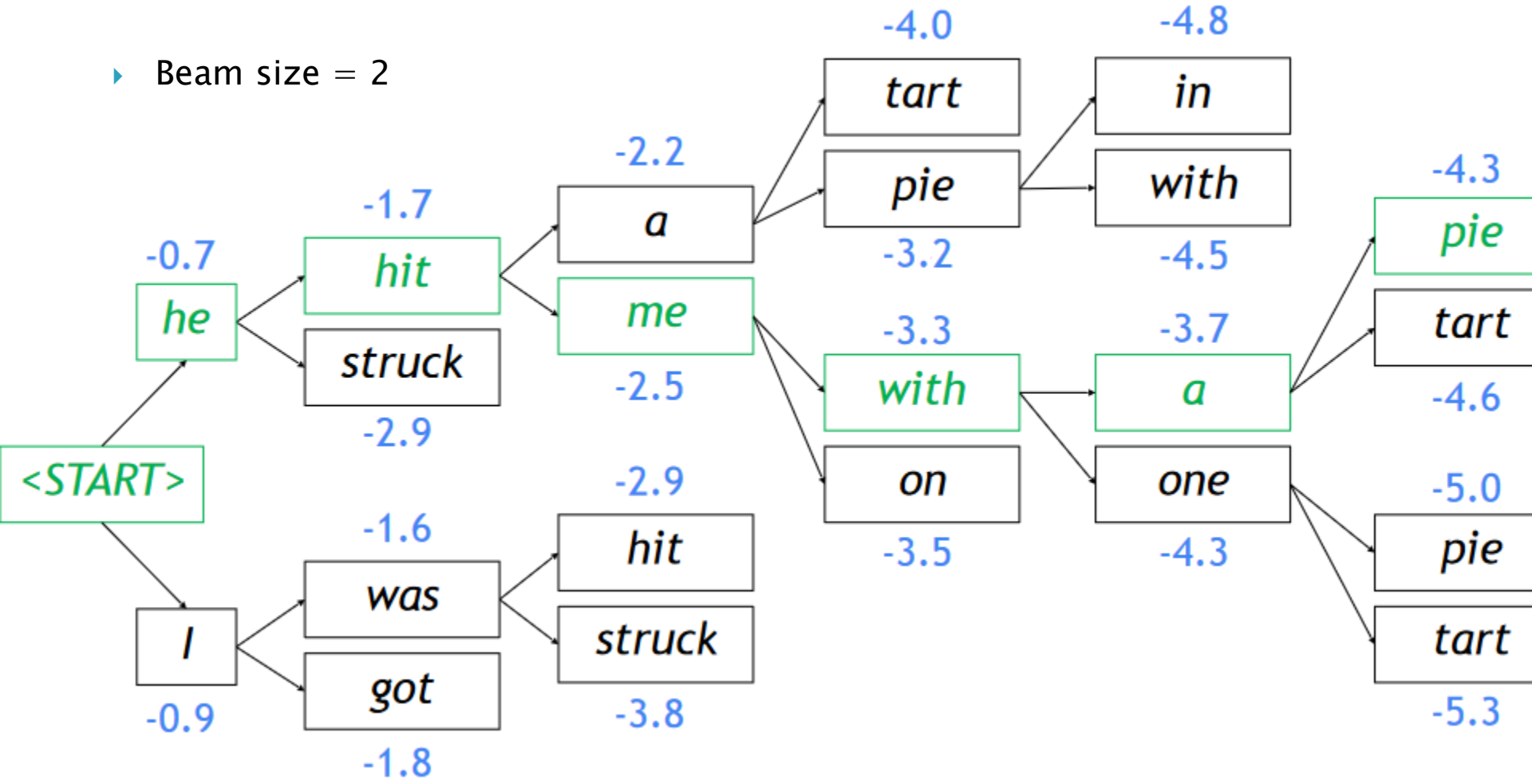
Beam search decoding

- Beam size = 2



Beam search decoding

- Beam size = 2



Generating new text with LM

- P(<s>) = 1
- P(A | <s>) = 1
- P(B |<s>,A) = 0.6
- P(</s>|<s>,A) = 0.4
- P(C | A , B) = 0.4
- P(</s>| A , B) = 0.6
- By greedy method, we got <s> A B </s>, but it is not the sentence with the highest probability.
- <s> A </s> is the sentence with the highest probability.

LM favors short sentences

- P(A | B , C) = 0.6
- P(</s>| B , C) = 0.4
- P(“ A B C ”) or P(“ A B C A B C A B C A B C ”) has a higher probability?

Reading list

- Speech and Language Processing, version 3
  - Chapter 3, 7 and 9