Natural Language Processing

Meeting 2024-11-20

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

Date	Activity
2024-09-04	Meeting 1
2024-10-02	Meeting 2
2024-11-20	Meeting 3
2025-01-15	Evaluation meeting
2025-03-20	Additional examination 1
2025-06-12	Additional examination 2

This session

18:00	Introduction & announcements
18:15	Transformer-based models: Q&A
18:30	Exercise: Attention
18:45	Introduction to lab 3
19:00	Break
19:15	Structured prediction: Q&A
19:30	Exercise: Feature engineering
19:45	Introduction to lab 4

Quiz 3.3, question 1

Suppose we encode the sentence “Gold is heavier than silver” using a bi-directional RNN. What issue does the recency bias cause?

1. The final hidden state contains more information about “Gold” than about “heavier”.

2. The final hidden state contains more information about “Gold" than about “silver”.

3. The final hidden state contains more information about “silver” than about “Gold”.

Quiz 3.3, question 1

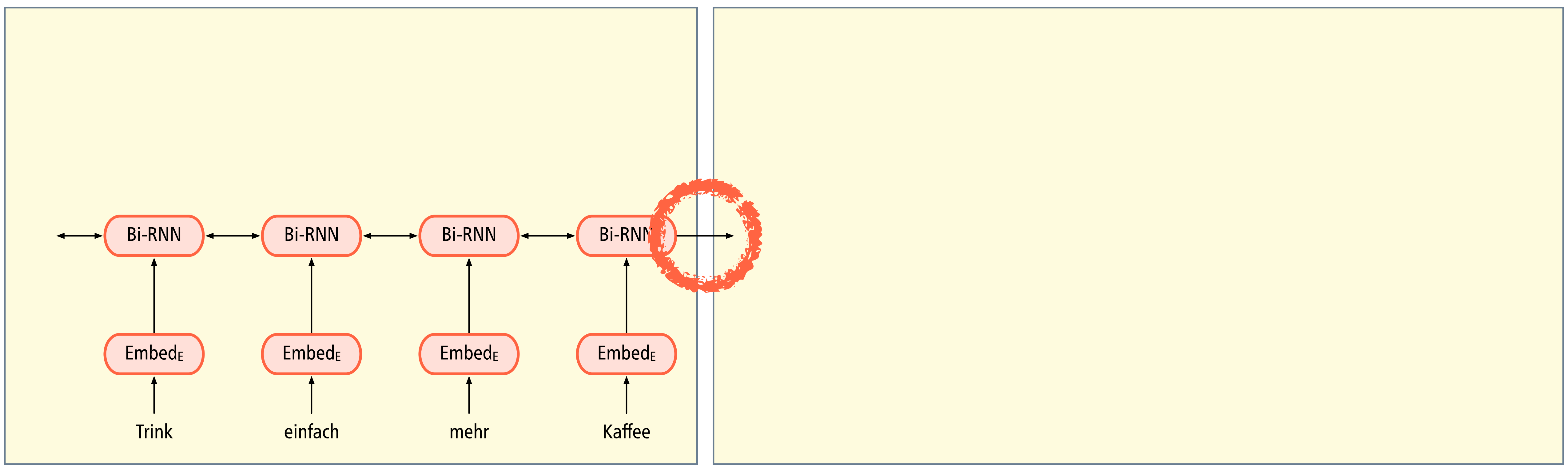
Suppose we encode the sentence “Gold is heavier than silver” using a bi-directional RNN. What issue does the recency bias cause?

1. The final hidden state contains more information about “Gold” than about “heavier”.

2. The final hidden state contains more information about “Gold" than about “silver”.

3. The final hidden state contains more information about “silver” than about “Gold”.

Recency bias in recurrent neural networks



encoder decoder

[Sutskever et al. (2014)]

Quiz 3.4, question 5

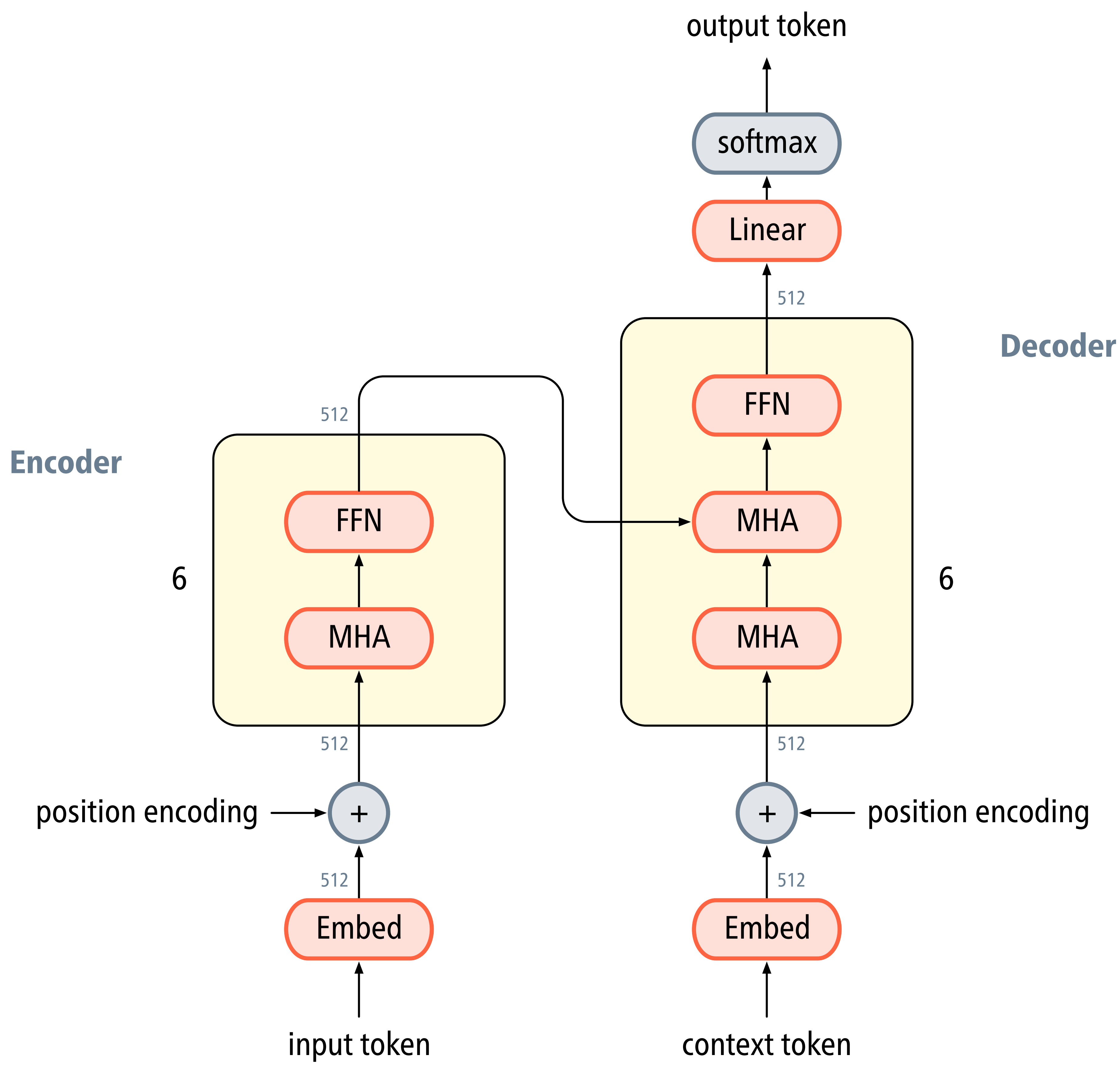
True or false: Permuting the input tokens to a Transformer encoder does not change the final token representations.

1. True
2. False
3. Depends on the position encodings

Quiz 3.4, question 5

True or false: Permuting the input tokens to a Transformer encoder does not change the final token representations.

1. True
2. False
3. Depends on the position encodings



output token
softmax
Linear
512
Decoder
FFN
MHA
MHA
6
512
Encoder
FFN
MHA
6
512
position encoding
+
512
Embed
input token
512
+
position encoding
512
Embed
context token

Quiz 3.3, question 3

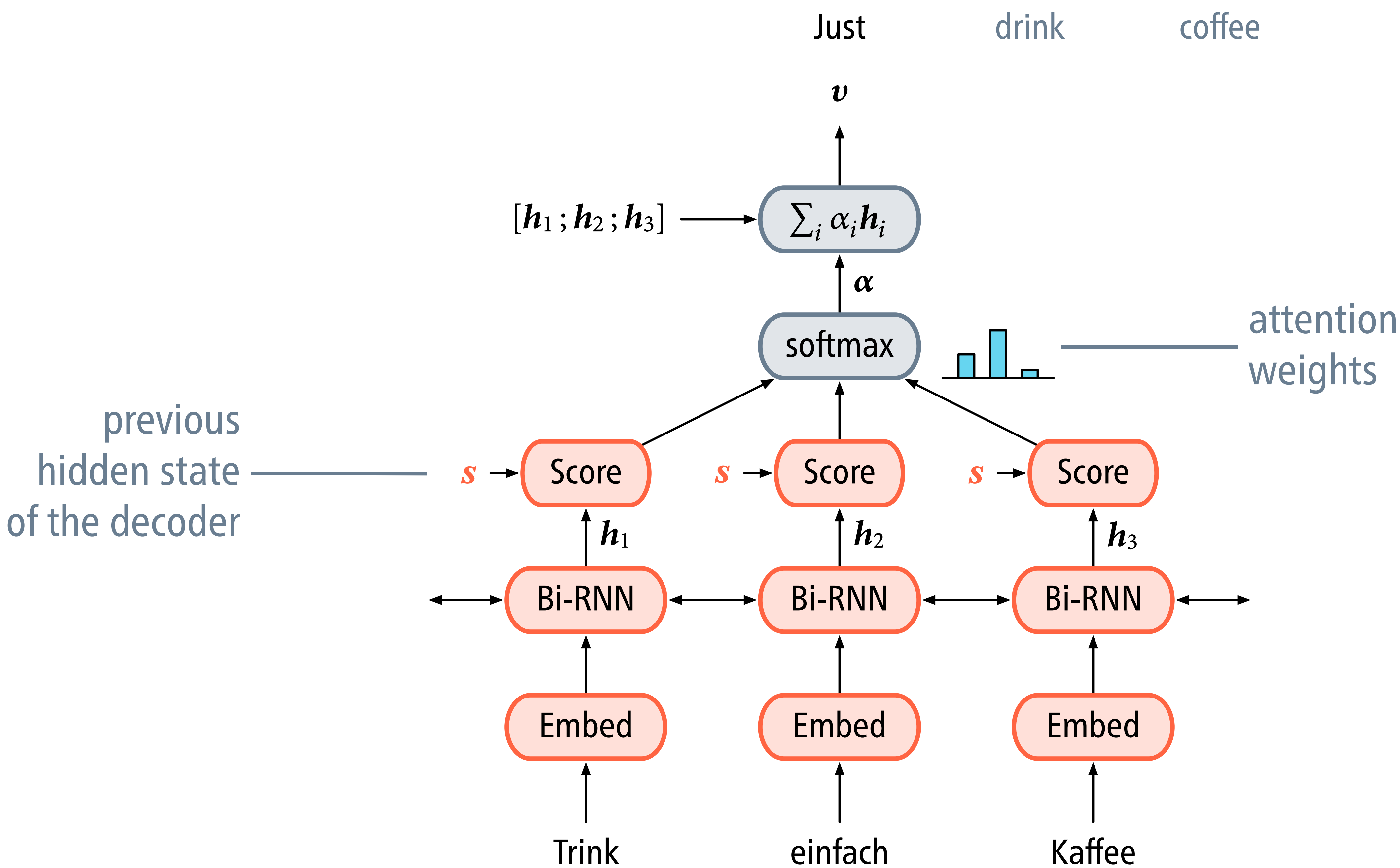
Consider the following values for the example in slide 6 ff.:

$\boldsymbol{s} = [0.4685, 0.9785]$ $\boldsymbol{h}_1 = [0.5539, 0.7239]$

$\boldsymbol{h}_2 = [0.4111, 0.3878]$ $\boldsymbol{h}_3 = [0.2376, 0.1264]$

Assuming that the attention score is computed using the dot product, what is $\boldsymbol{v}$?

Attention for translation



Bahdanau et al. (2015)

Quiz 4.3, question 2

In the central matrix for the run of the Viterbi algorithm on the example, what is the intended value in row VERB, column “see”?

1. the maximal score when tagging the first 4 words in the sentence and where see is tagged as a VERB

2. the maximal score when tagging the first 4 words in the sentence as DET NOUN AUX VERB

3. the maximal score when tagging the word see as a VERB

Quiz 4.3, question 2

In the central matrix for the run of the Viterbi algorithm on the example, what is the intended value in row VERB, column "see"?

1. the maximal score when tagging the first 4 words in the sentence and where see is tagged as a VERB

2. the maximal score when tagging the first 4 words in the sentence as DET NOUN AUX VERB

3. the maximal score when tagging the word see as a VERB

		the$_1$	man$_2$	can$_3$	see$_4$	a$_5$	can$_6$	
[BOS]	−0.00							
ADJ		−17.29	−18.55	−28.99	−32.49	−39.28	−40.87	
AUX		−17.72	−23.37	−16.23	−32.81	−36.63	−34.72	
DET		−3.00	−21.90	−29.70	−33.03	−25.31	−44.21	
NOUN		−17.85	−10.46	−22.09	−35.70	−33.72	−35.38	
PROPN		−16.25	−19.49	−29.34	−35.38	−39.91	−41.80	
VERB		−17.47	−21.62	−27.97	−22.09	−40.88	−43.93	
[EOS]								−39.50

The central invariant

- We ensure that the value in row t, column i is the maximal score that can be obtained when tagging the first i words in the sentence in such a way that word number i is tagged as t.

  column = maximal scores, sub-categorised by last tag

- If the algorithm can establish and maintain this invariant, then we can read off the maximal score for the complete sentence from the last column (end-of-sentence).

Quiz 4.4, question 3

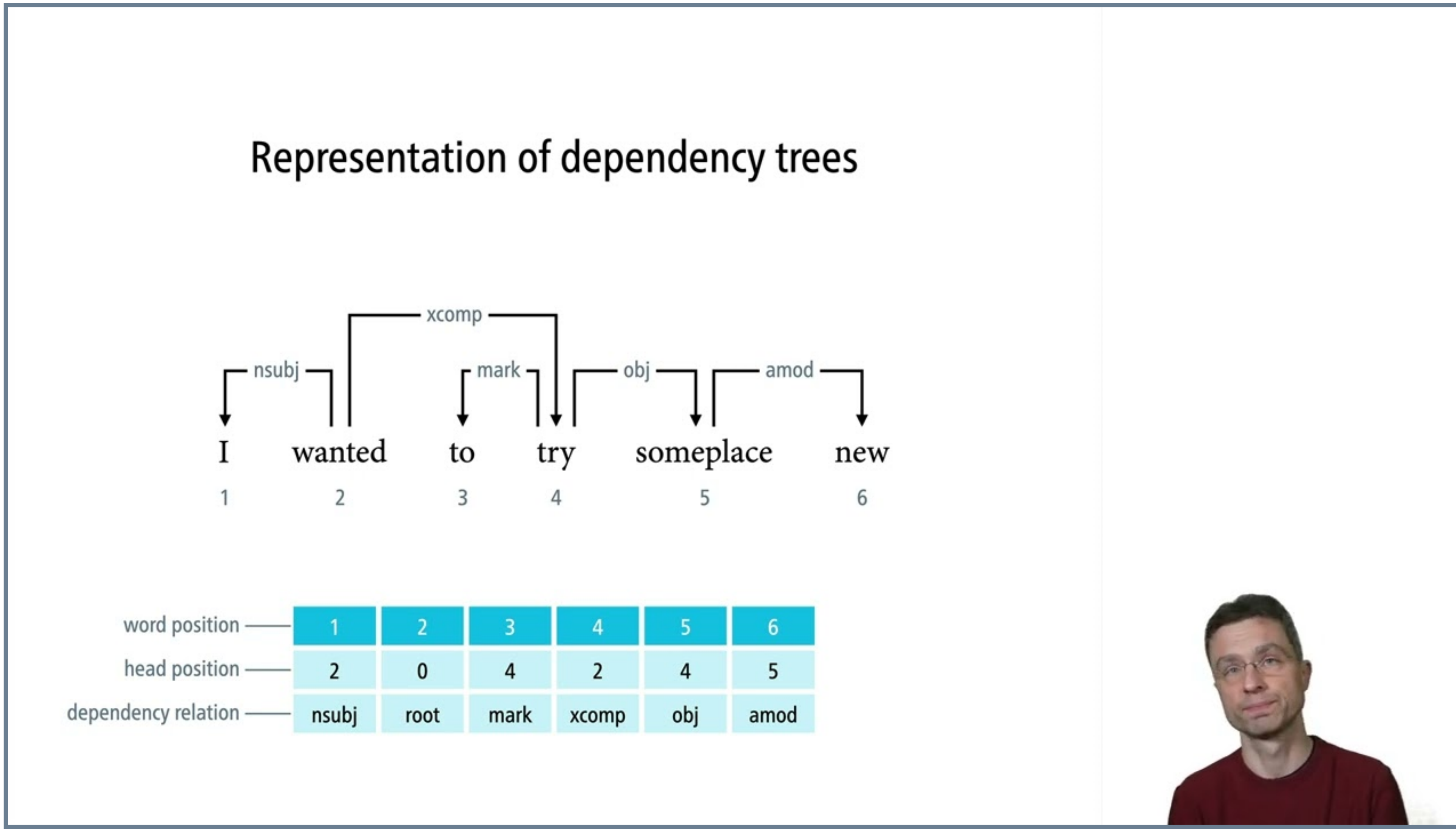
We are interested in the number of potential dependency trees for a sentence consisting of 20 words. How many digits are in that number?

1. 15
2. 25
3. 35

Quiz 4.4, question 3

We are interested in the number of potential dependency trees for a sentence consisting of 20 words. How many digits are in that number?

1. 15
2. 25
3. 35



$$20^{20-1} = 52428\ 80000\ 00000\ 00000\ 00000$$

Quiz 4.4, question 4

We implement a dependency parser based on the Chiu–Liu/Edmonds algorithm. By which factor would we expect the parser runtime to grow when we double the sentence length?

1. 2
2. 3
3. 8

Quiz 4.4, question 4

We implement a dependency parser based on the Chiu–Liu/Edmonds algorithm. By which factor would we expect the parser runtime to grow when we double the sentence length?

1. 2
2. 3
3. 8

Computational complexity

- Under the arc-factored model, the highest-scoring dependency tree can be found in $O(n^3)$ time (n = sentence length).

  Chu–Liu/Edmonds algorithm; McDonald et al. (2005)

- Even seemingly minor extensions of the arc-factored model entail intractable parsing.

  McDonald and Satta (2007)

- For some of these extensions, polynomial-time parsing is possible for restricted classes of dependency trees.

Quiz 4.4, question 5

In the context of dependency parsing, an “oracle” is a function …

1. … that maps a dependency tree
   to a sequence of transitions.

2. … that maps a sentence
   to its gold-standard dependency tree.

3. … that maps a sequence of transitions
   to a dependency tree.

Quiz 4.4, question 5

In the context of dependency parsing, an "oracle" is a function …

1. … that maps a dependency tree
   to a gold-standard sequence of transitions.

2. … that maps a sentence
   to its gold-standard dependency tree.

3. … that maps a sequence of transitions
   to a dependency tree.

Training transition-based dependency parsers

- To train a transition-based dependency parser, we need a treebank with gold-standard dependency trees.
- In addition to that, we need an algorithm that tells us the gold-standard transition sequence for a tree in that treebank.
- Such an algorithm is conventionally called an **oracle**.

Quiz 4.5, question 1

Here are three dependency trees on three-word sentences. Each tree is specified by the list of the head positions, as introduced in Lecture 4.4. Which of these trees is not projective?

1. 2 0 2
2. 2 0 1
3. 0 3 1

Quiz 4.5, question 1

Here are three dependency trees on three-word sentences. Each tree is specified by the list of the head positions, as introduced in Lecture 4.4. Which of these trees is not projective?

1. 2 0 2
2. 2 0 1
3. 0 3 1

Quiz 4.5, question 4

Which of the following is true of the arc-standard algorithm?

1. Every valid transition sequence
   builds some dependency tree.

2. Every dependency tree
   can be built by some valid transition sequence.

3. Every valid transition sequence
   builds the correct dependency tree.

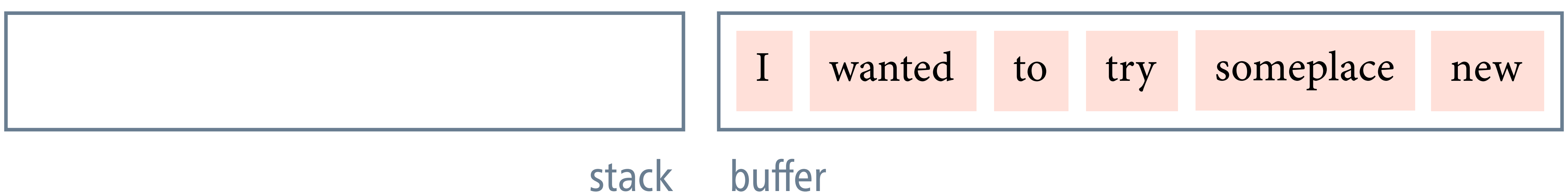
Quiz 4.5, question 4

Which of the following is true of the arc-standard algorithm?

1. Every valid transition sequence builds some dependency tree.
2. Every dependency tree can be built by some valid transition sequence.
3. Every valid transition sequence builds the correct dependency tree.

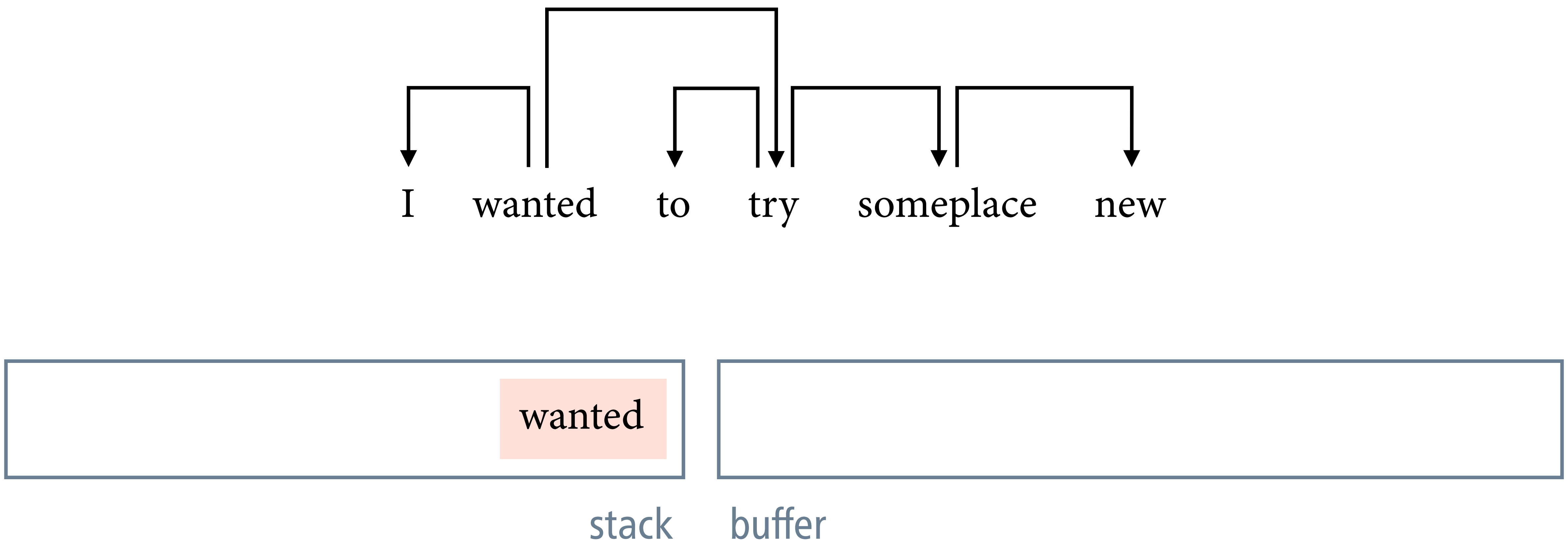
Example run

I wanted to try someplace new



(initial configuration)

Example run



(terminal configuration)

Soundness and completeness

- **Soundness**

  Every valid transition sequence that starts in the initial configuration and ends in some terminal configuration builds some projective dependency tree.

- **Completeness**

  Every projective dependency tree can be built by some valid transition sequence that starts in the initial configuration and ends in some terminal configuration.

Quiz 4.6, question 2

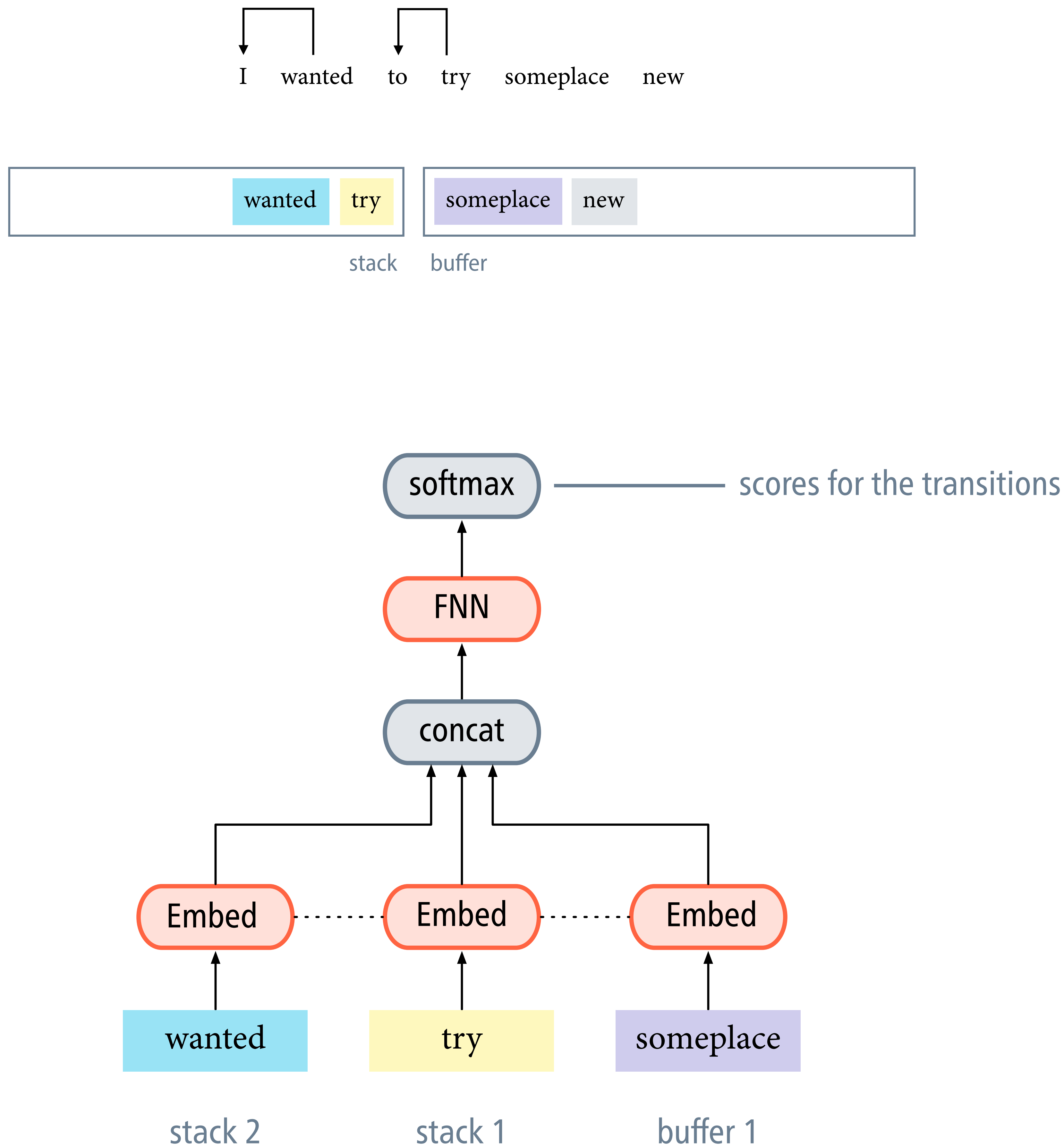
Consider the second configuration from the Chen and Manning (2014) example, and suppose that in this configuration, the parser makes the transitions SH RA. In the new configuration, what is the value of the stack 1 feature?

1. wanted
2. try
3. someplace

Quiz 4.6, question 2

Consider the second configuration from the Chen and Manning (2014) example, and suppose that in this configuration, the parser makes the transitions SH RA. In the new configuration, what is the value of the stack 1 feature?

1. wanted
2. try
3. someplace


I
wanted
to
try
someplace
new
wanted
try
someplace
new
stack
buffer
softmax
scores for the transitions
FNN
concat
Embed
Embed
Embed
wanted
try
someplace
stack 2
stack 1
buffer 1