Natural Language Processing. Transformers. Part 4
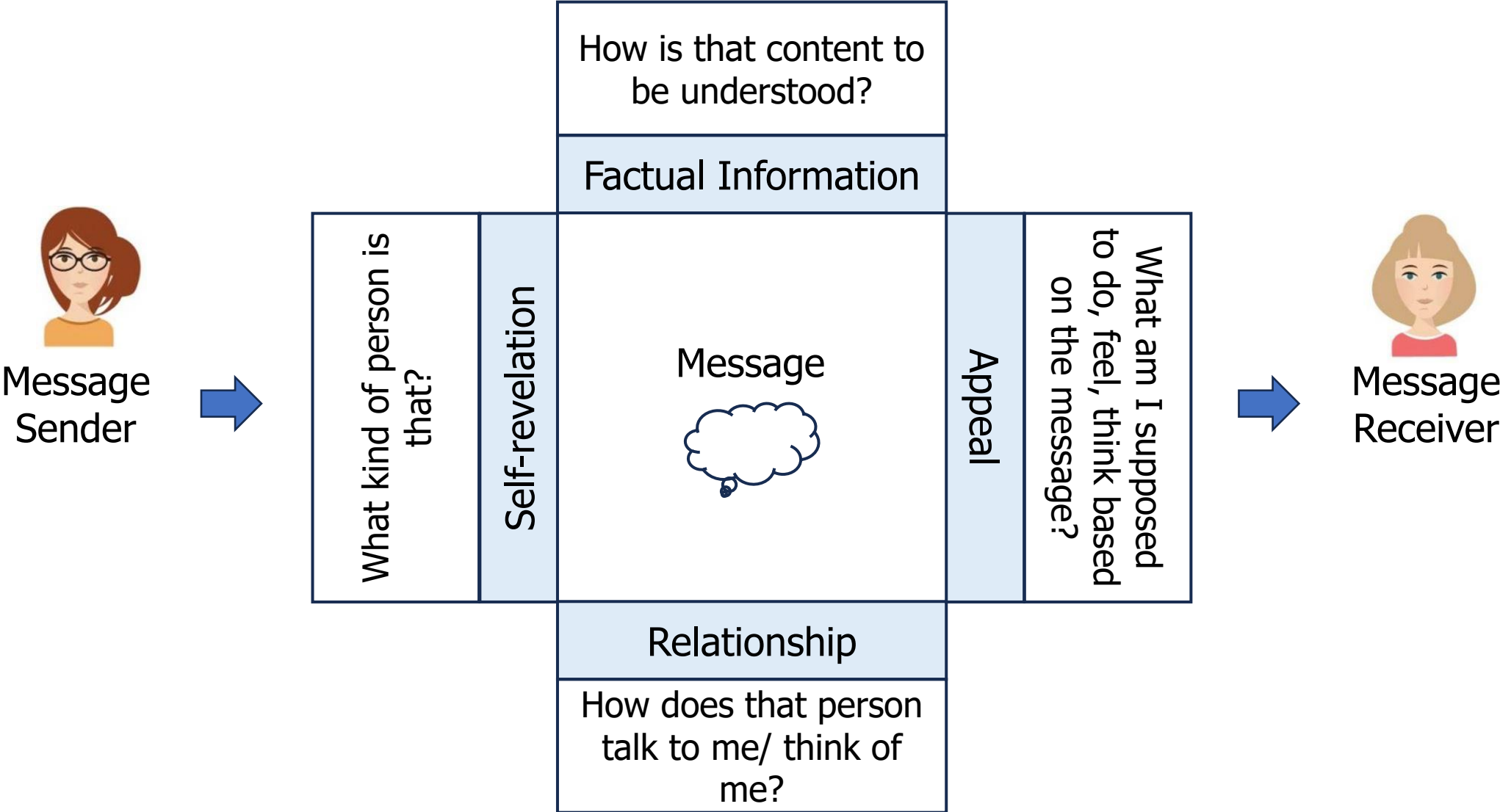## (A Very Short Introduction)

Sergey V. Axyonov,

PhD, Associate Professor,

Department of Fundamental Computer Science, Institute of Applied Math & Computer Science,

Tomsk State University

Tomsk-2024

Summary from the Previous Presentation

Word	living being	feline	human	gender	royalty	verb	plural
cat	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
kitten	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
dog	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
houses	-0.8	-0.4	-0.5	0.1	-0.9	-0.3	0.8
man	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
woman	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
king	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
queen	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

Word   Embedding Vector

We can code all text tokens with very long vectors representing different contexts in which those words occur.

Words with similar meanings have similar vectors

So we can get something like this «Queen – Woman = King – Man»

Four Aspects of a Message

Message Sender

What kind of person is that?

Self-revelation

How is that content to be understood?

Factual Information

Message

Relationship

How does that person talk to me/ think of me?

Appeal

What am I supposed to do, feel, think based on the message?

Message Receiver

An Example

Imagine the following situation: She sits at the wheel of the common car, he in the passenger seat. Now he says, "The traffic light's green." What can be heard depending on the quality of the relationship and the experiences made between the two protagonists so far?

- Factual information: The traffic light is green. We can drive!
- Self-revelation: I am much more qualified than you to drive a car, because I have already noticed that the traffic light is green!
- Relationship: I always have to tell you what to do!
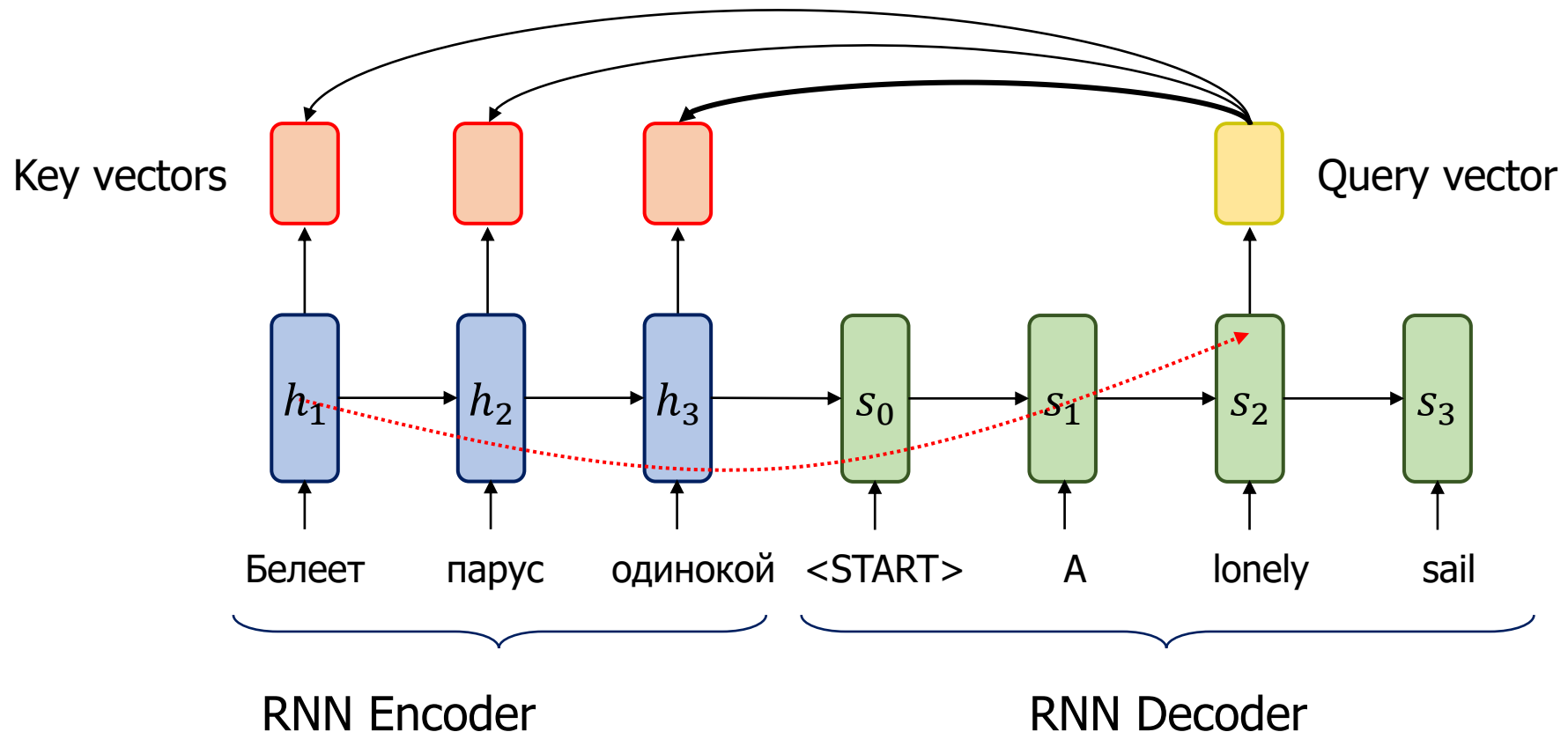- Appeal: Just drive!

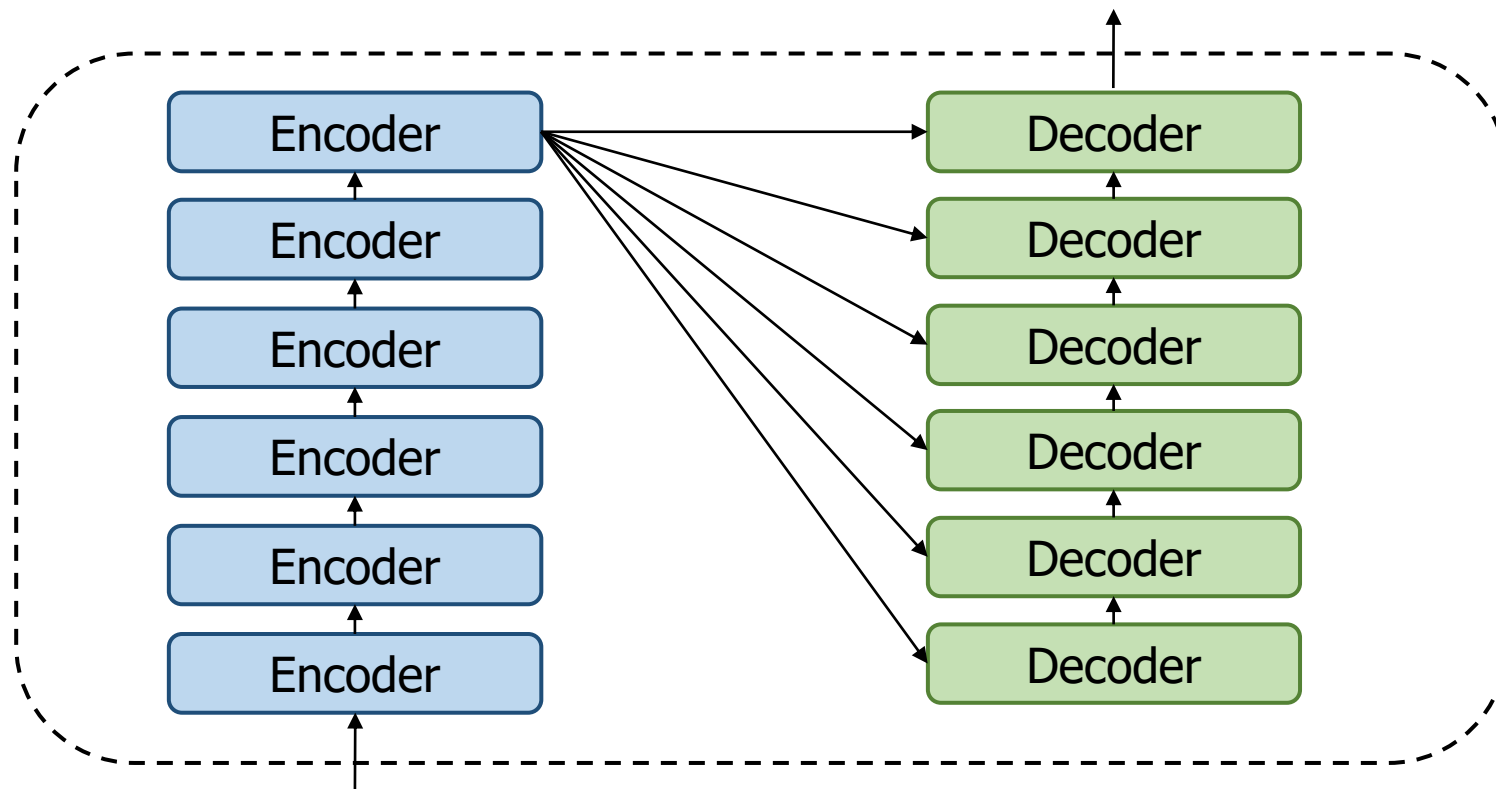# Seq2Seq2 Model
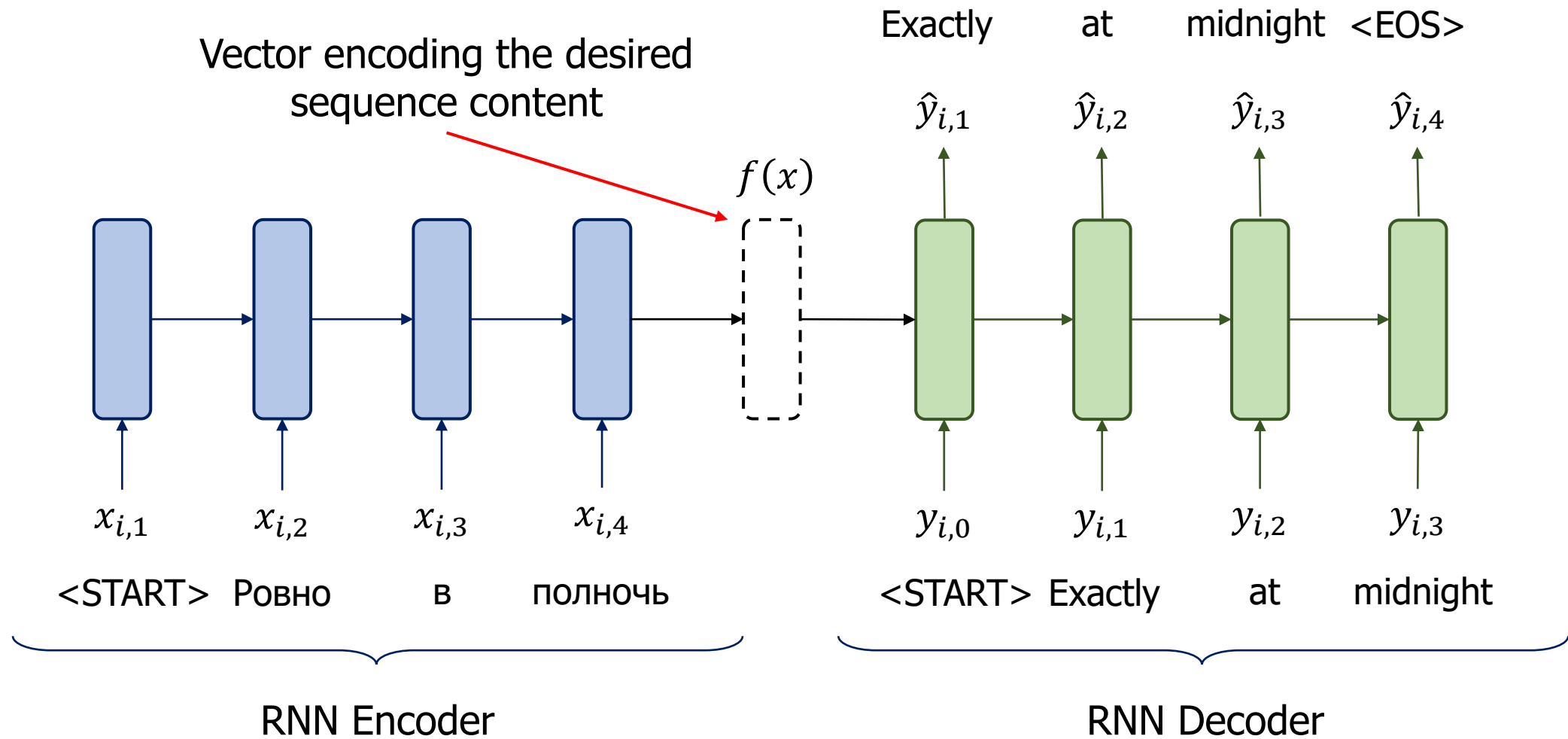
Missing Context
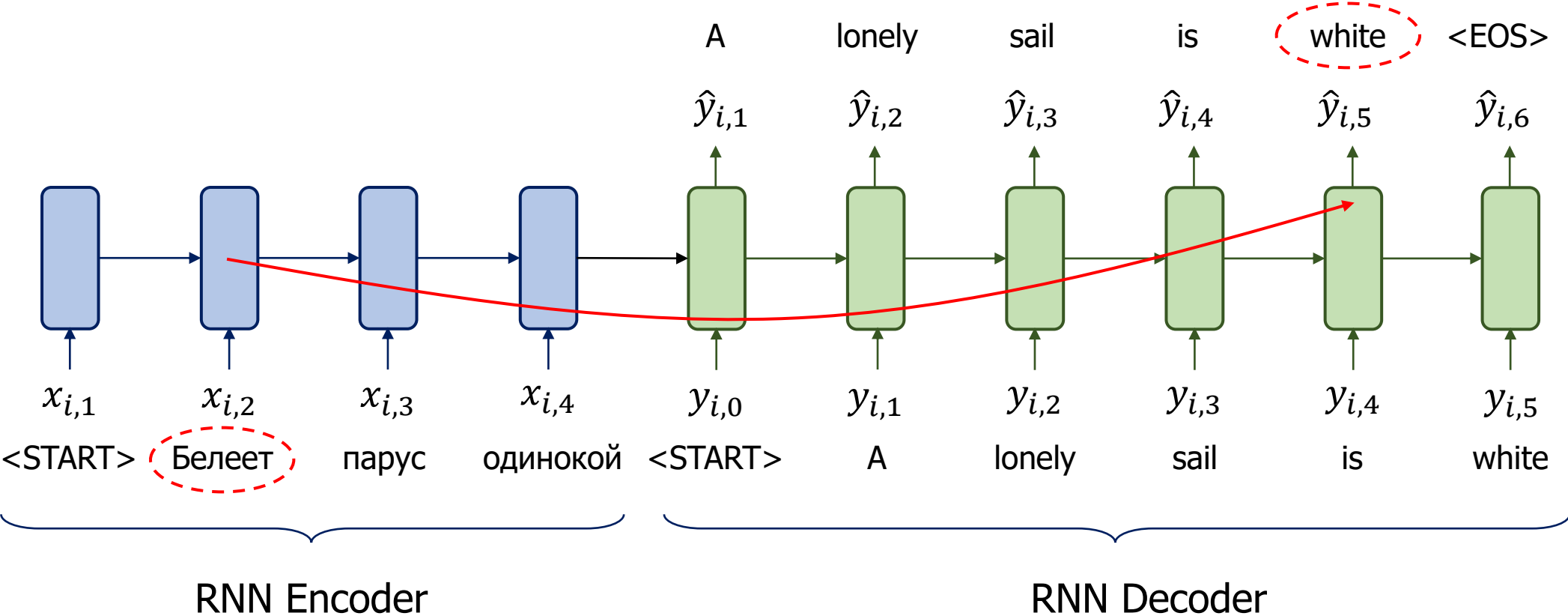
Attention Mechanism

Attention assigns importance to each word by calculating "soft" weights for the word's numerical representation within context window to determine its importance.

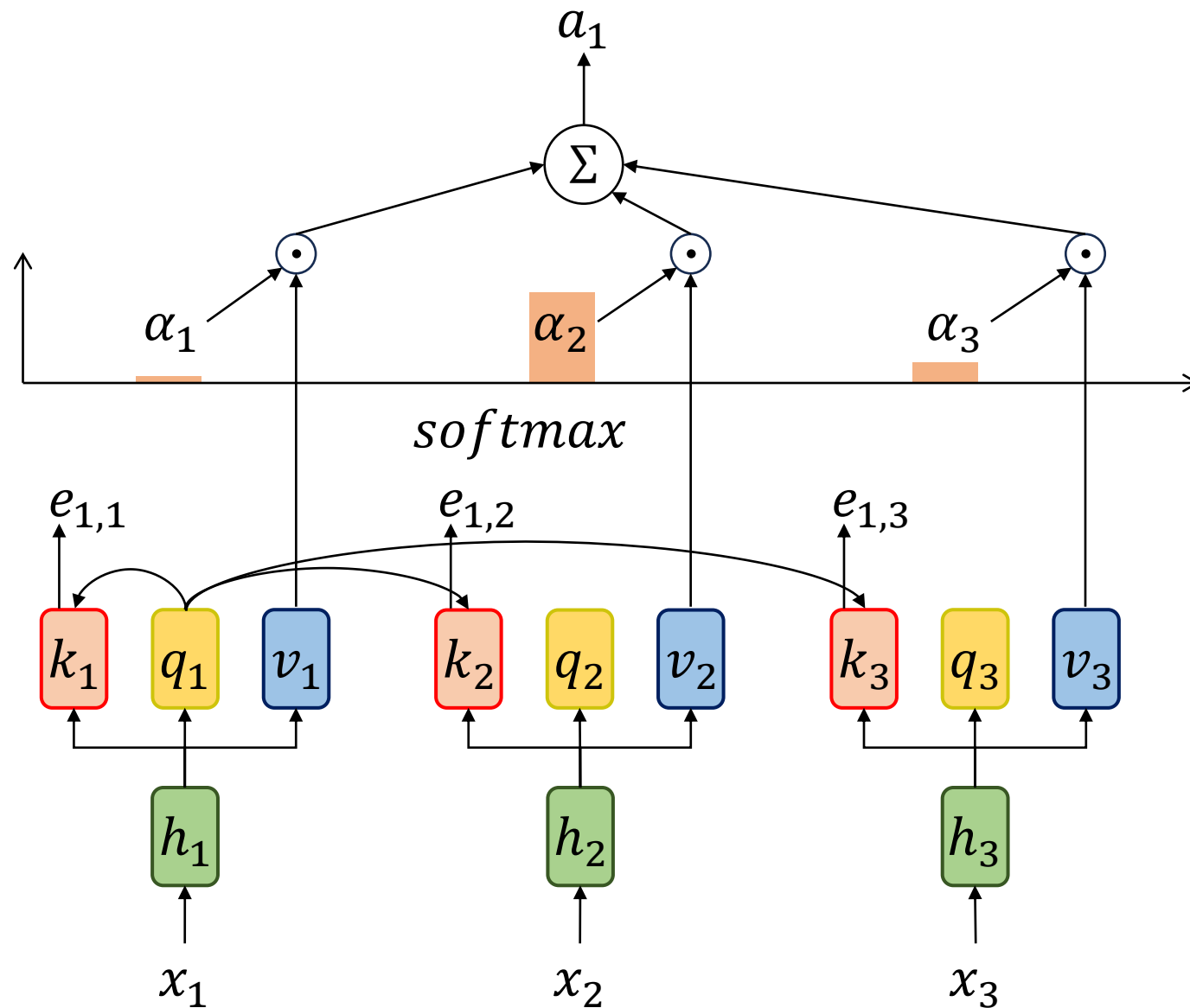Compare the query against each key to find the best match

Transformer General Structure

«Быть или не быть, вот в чём вопрос»
(«Гамлет», Уильям Шекспир)

Encoder
Encoder
Encoder
Encoder
Encoder
Encoder

Decoder
Decoder
Decoder
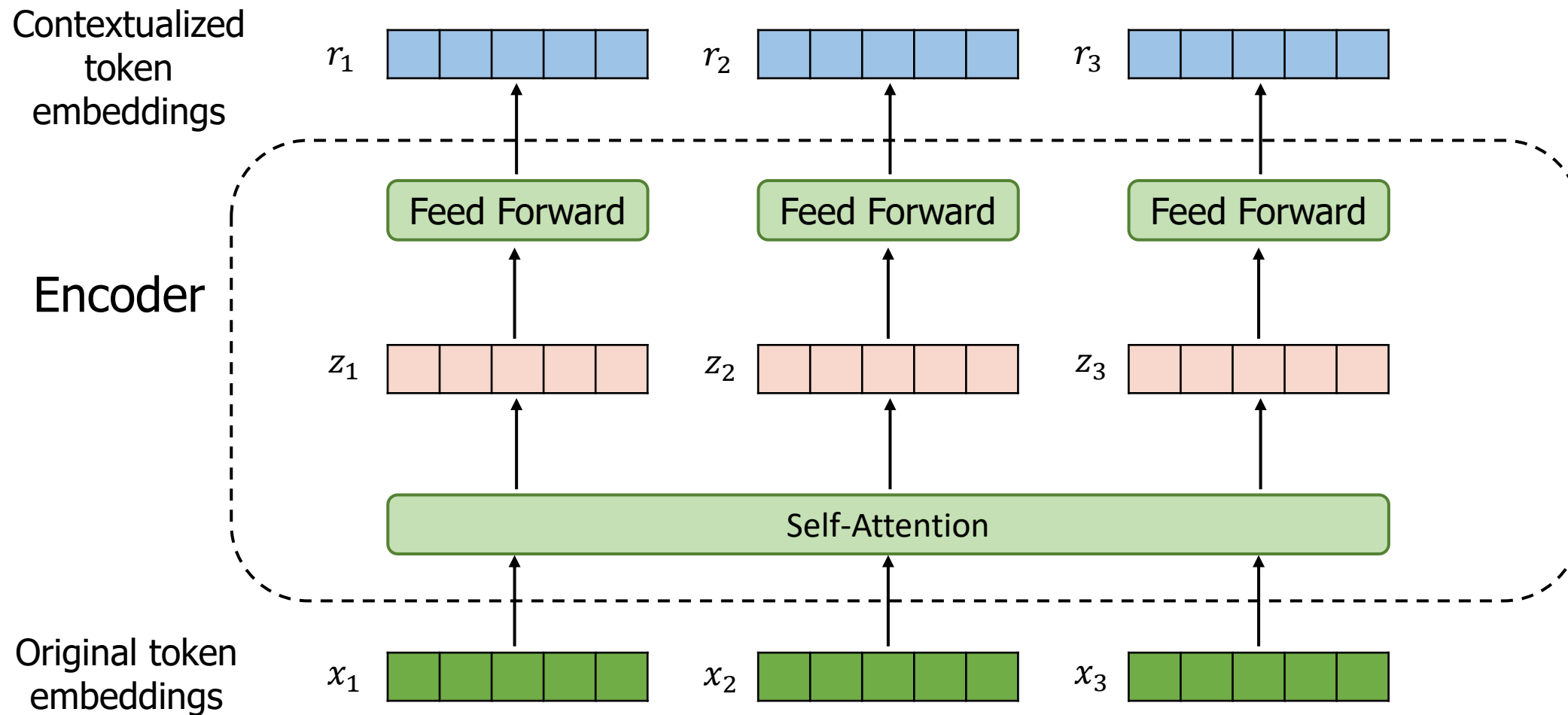Decoder
Decoder
Decoder

«To be, or not to be, that is the question»
(«Hamlet», William Shakespeare)

Self-Attention

Self-attention is a mechanism used in neural networks to capture dependencies and relationships within input sequences.

Self-Attention

Contextualized token embeddings

r_1 r_2 r_3

Encoder

Feed Forward   Feed Forward   Feed Forward

z_1 z_2 z_3

Self-Attention

Original token embeddings

x_1 x_2 x_3

Self-Attention: Example

«Mockingbirds don't do one thing but make music for us to enjoy. They don't eat up people's gardens, don't nest in corncribs, they don't do one thing but sing their hearts out for us. That's why it's a sin to kill a mockingbird.» («To Kill a Mockingbird», Harper Lee)

| Mockingbirds give us only good things	The music of these birds gives us pleasure		These birds should be protected	
Mockingbirds try to sing well	The music of these birds is nice	They do their best	They are good singing birds.	They don't do anything wrong
Mockingbirds are singing birds	Music made by mockingbirds	birds sing their heart out	They = Mockingbirds	Mockingbirds don't eat up gardens
mockingbirds	music	heart out	they	don't eat

Some More Examples
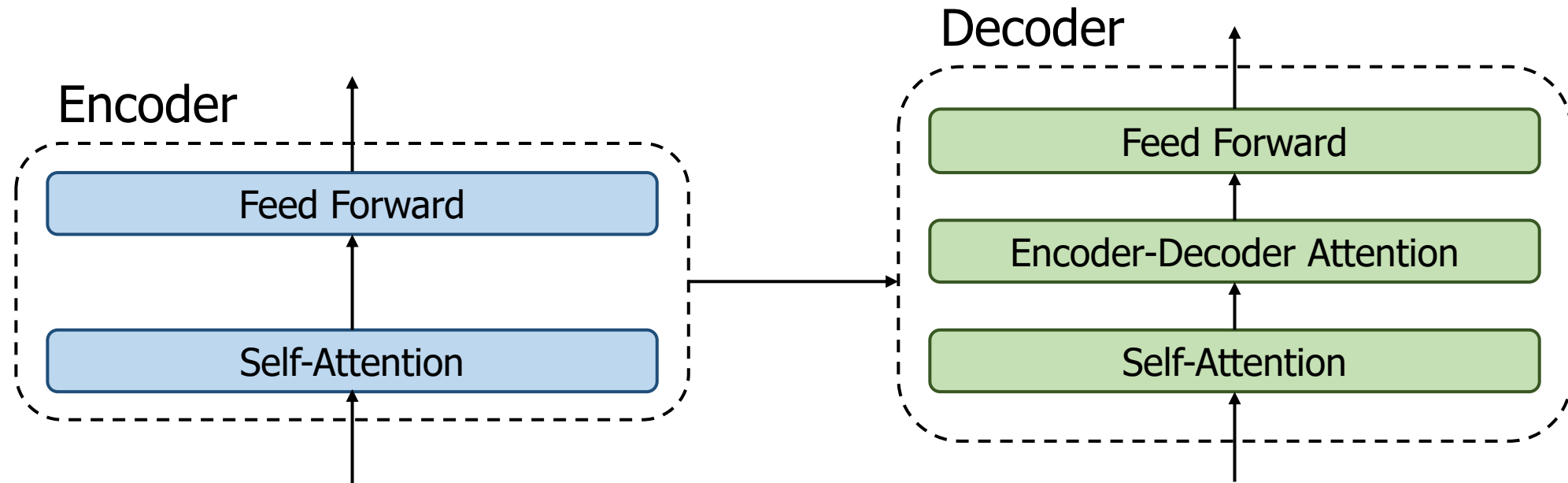
1.What word can you make shorter by adding two letters? Short.

2.Police arrested two kids yesterday, one was drinking battery acid, the other was eating fireworks. They charged one – and let the other one off.

3.What would the Terminator be called in his retirement? The Exterminator.

4.Sundays are always a little sad, but the day before is a sadder day.

5.I lost my job at the bank on my first day. A woman asked me to check her balance, so I pushed her over.

6.How much money does a skunk have? Only one scent.

7.Why is Peter Pan always flying? Because he Neverlands.

Multi-Head Attention

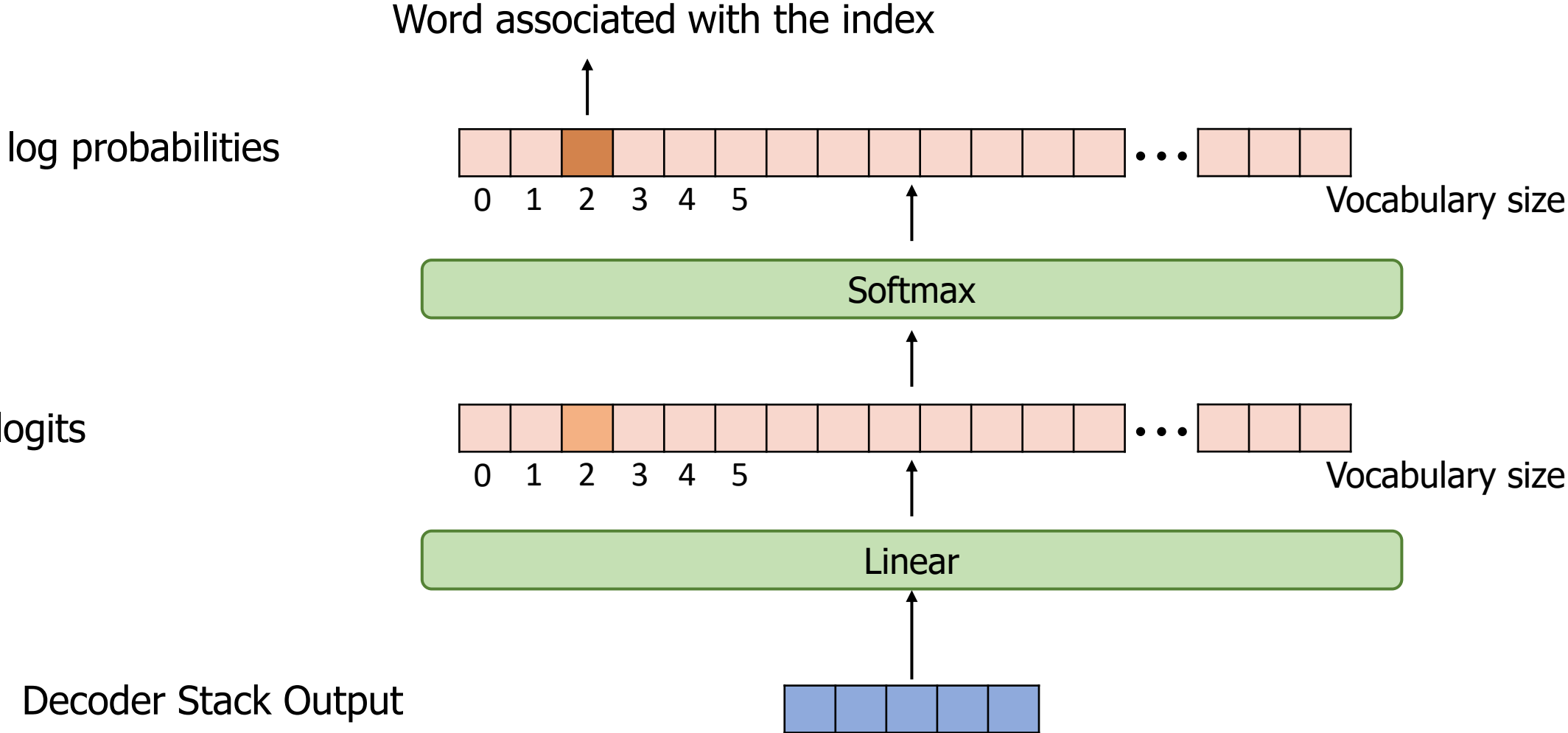$$a_2 = \begin{bmatrix} a_{2,1} \\ a_{2,2} \\ a_{2,3} \end{bmatrix}$$
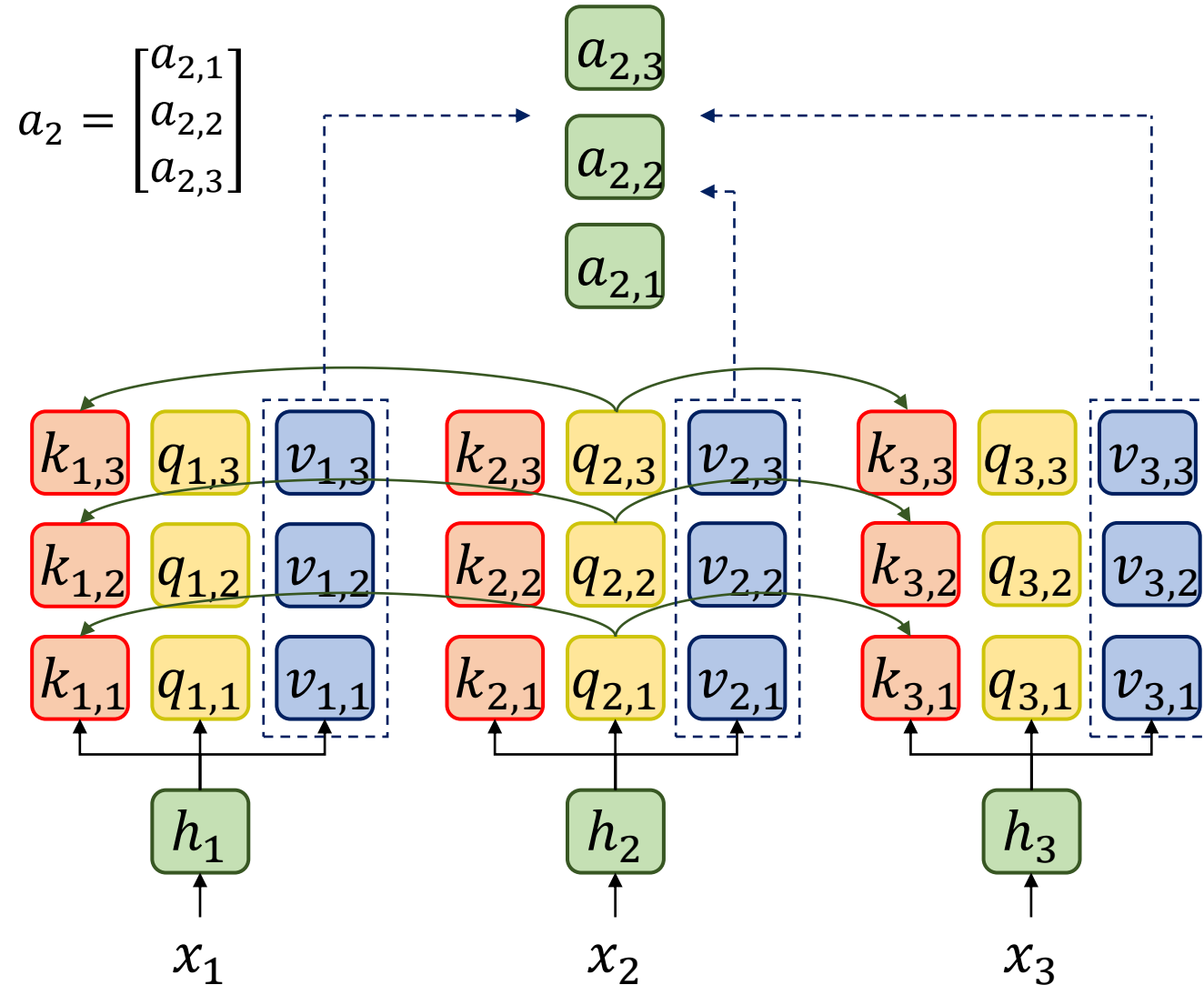
- ✓ Ability to capture diverse aspects of relationships between tokens
- ✓ Multiple key, query and value vectors.
- ✓ Training weights independently for each attention head
- ✓ In practice, 8 attention heads give a good result.

Encoder/Decoder Attention

Final Predictions

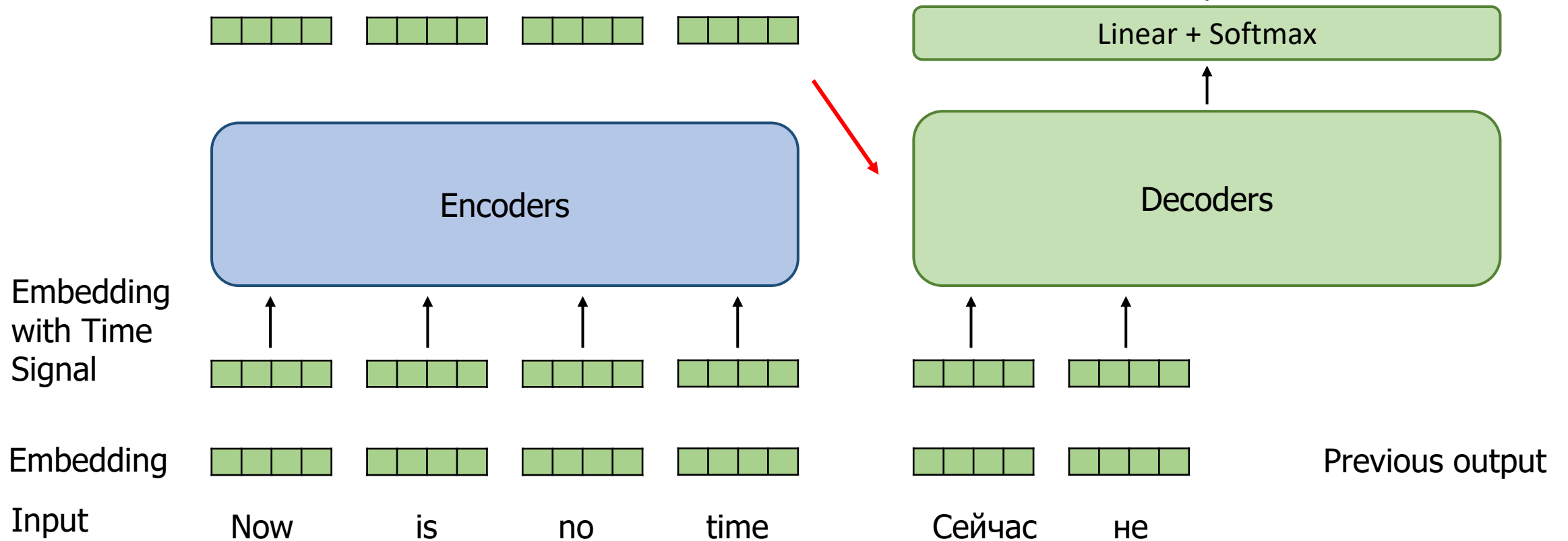Word associated with the index

log probabilities

| | | | | | | | | | | | | | | | ... | | | |

0 1 2 3 4 5

Vocabulary size

Softmax

logits

| | | | | | | | | | | | | | | | ... | | | |

0 1 2 3 4 5

Vocabulary size

Linear

Decoder Stack Output

Getting a New Sequence

Decoding time step: 3

Сейчас   не   время

Embedding
with Time
Signal

Embedding   Previous output

Input   Now   is   no   time   Сейчас   не

Linear + Softmax

Encoders   Decoders

"Now is no time to think of what you do not have. Think of what you can do with what there is."
(«The Old Man and the Sea», Ernest Hemingway)

Text Generation