Внимание и трансформеры

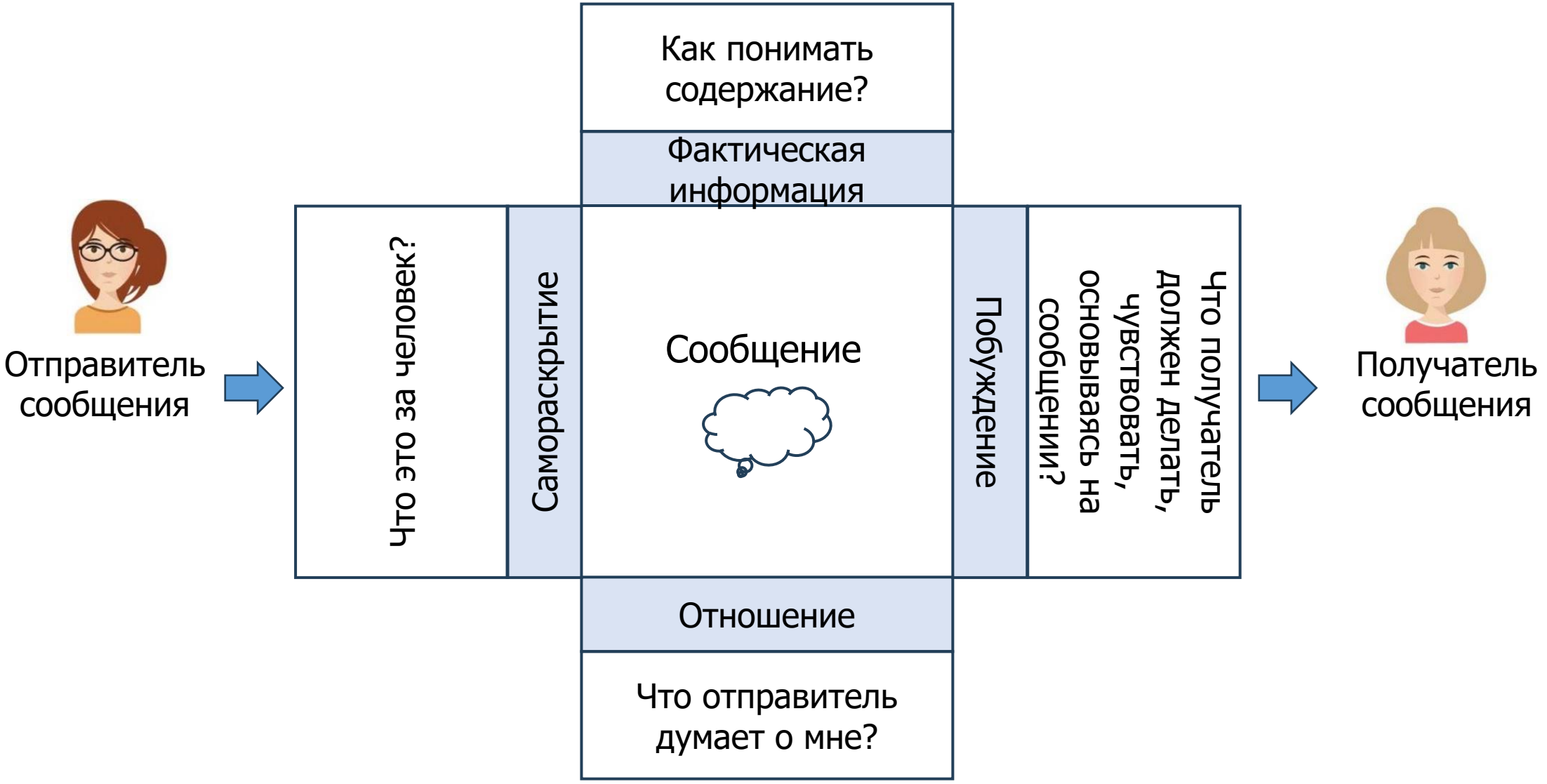
Сергей В. Аксёнов,

к.т.н., доцент каф. Автоматизации обработки информации,

Томский университет систем управления и радиоэлектроники

Томск-2024

Аспекты сообщения

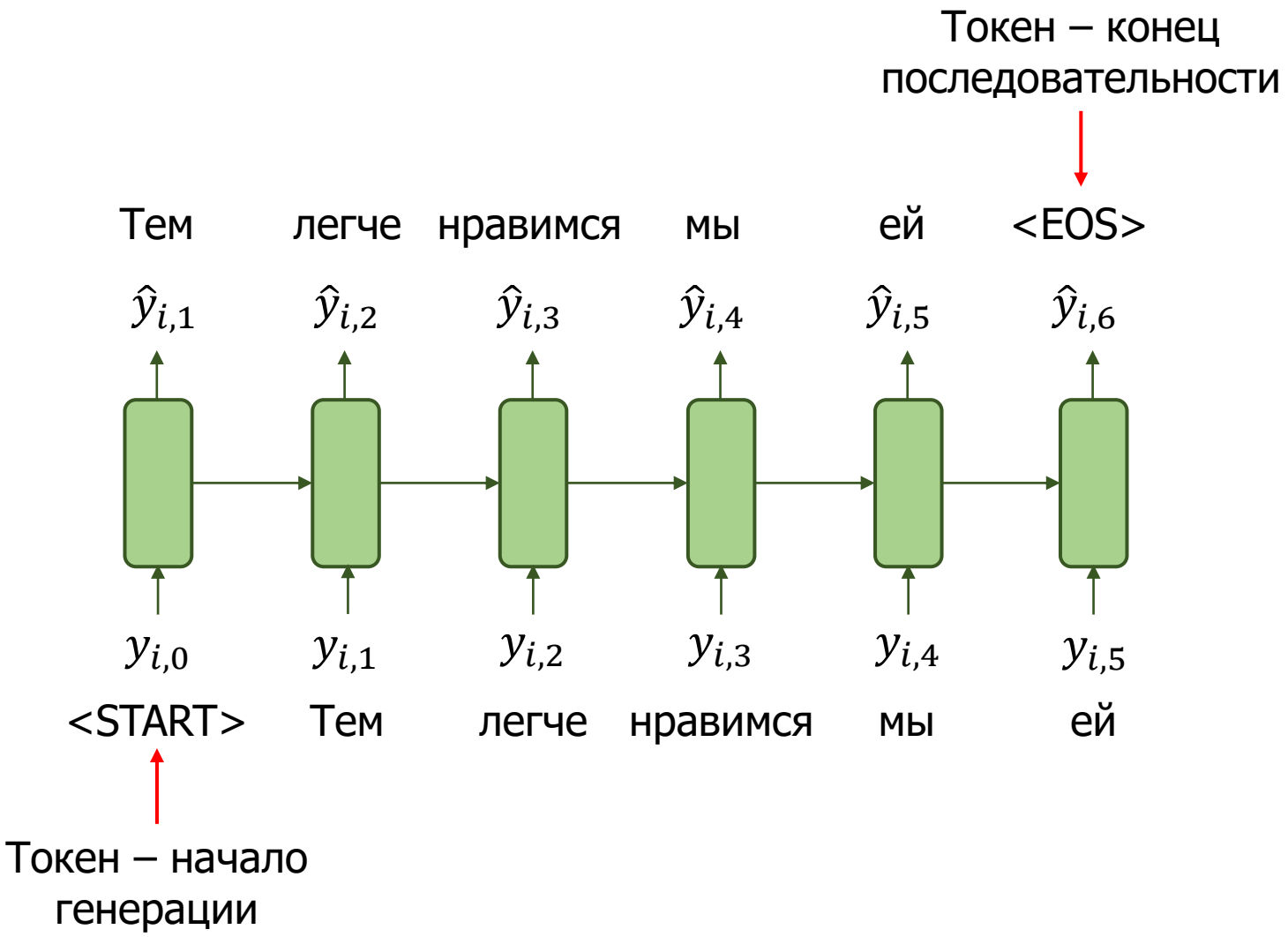


Пример

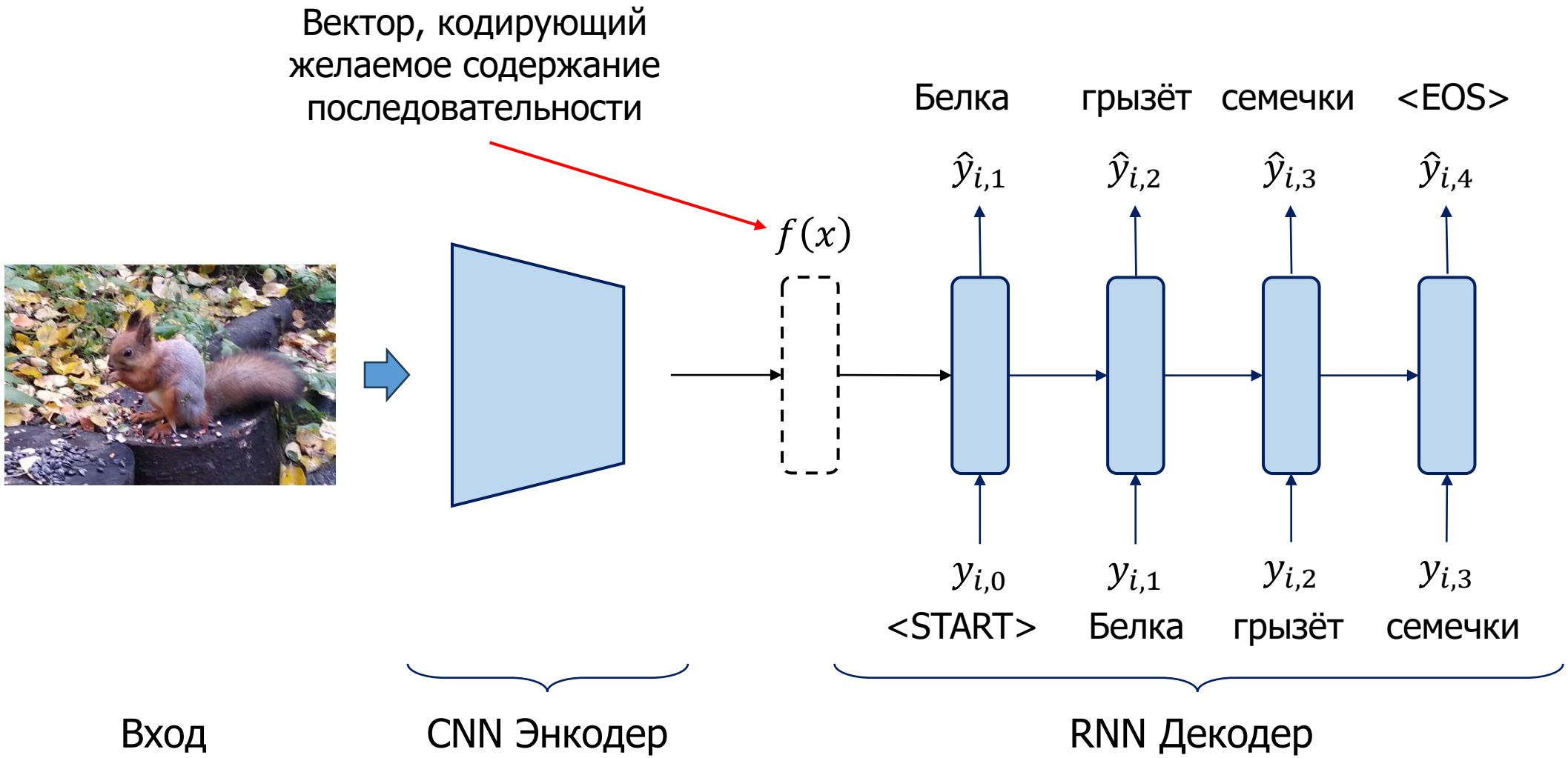
Представьте себе следующую ситуацию: девушка сидит за рулем обычной машины, парень - на пассажирском сиденье. Теперь он говорит: «Светофор зеленый». Что можно услышать в зависимости от качества отношений и опыта, полученного между двумя главными героями на данный момент?

- • Фактическая информация: Светофор – зелёный. Мы можем ехать!
- • Самораскрытие: Я гораздо более квалифицирован, чем вы, в управлении автомобилем, потому что уже заметил, что светофор зеленый!
- • Отношение: Я всегда должен говорить тебе, что делать!
- • Побуждение: Поехали!

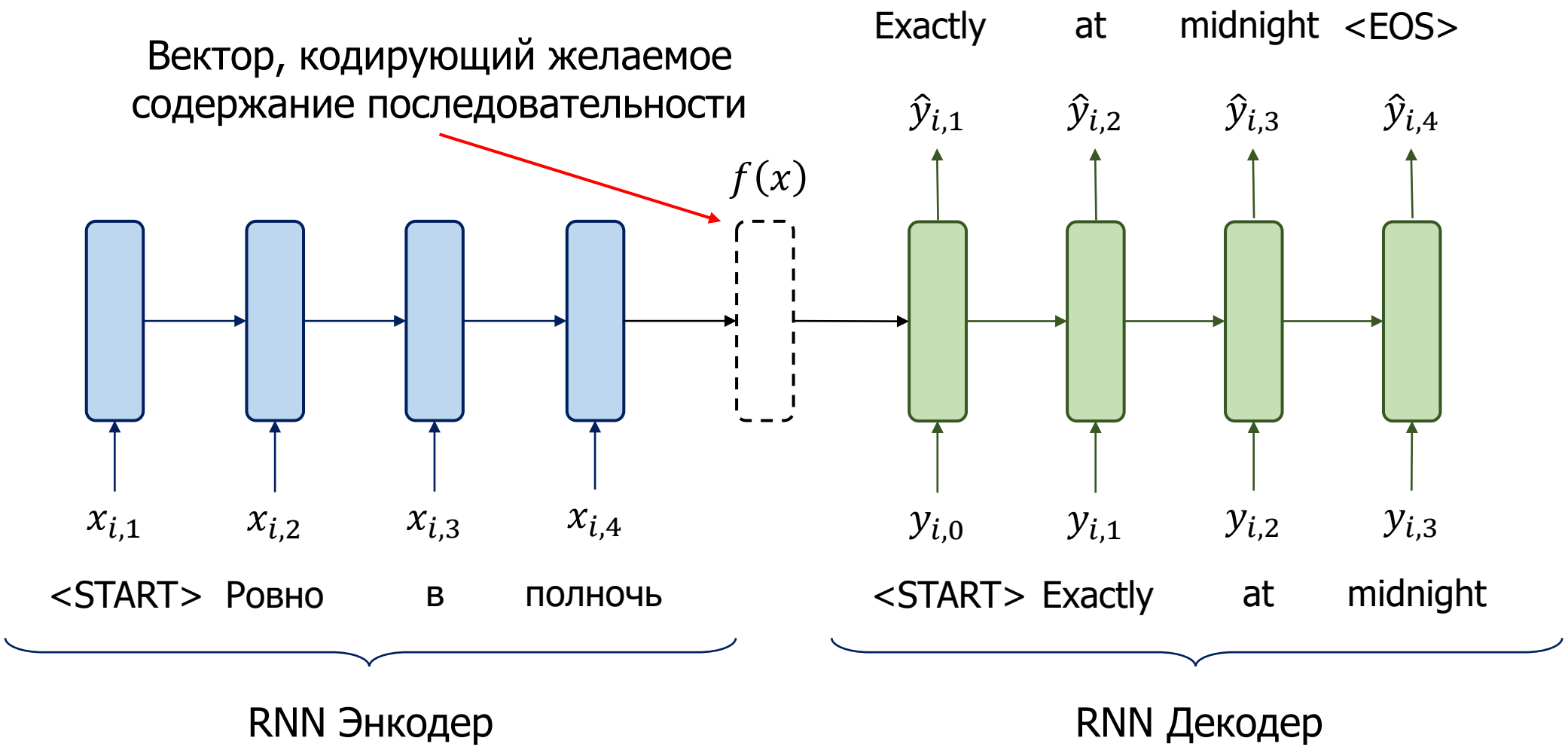
Простая нейросетевая языковая модель



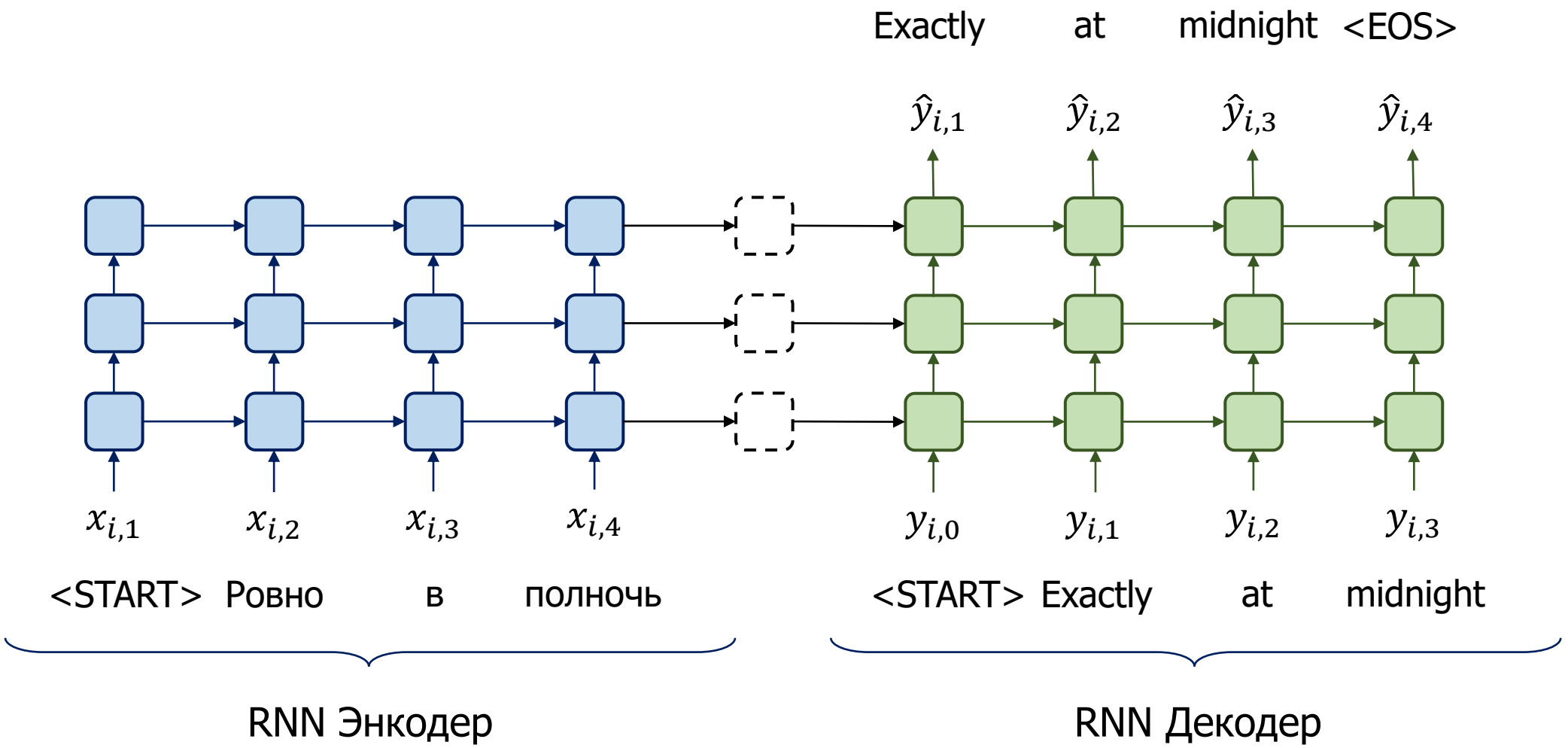
Условная языковая модель



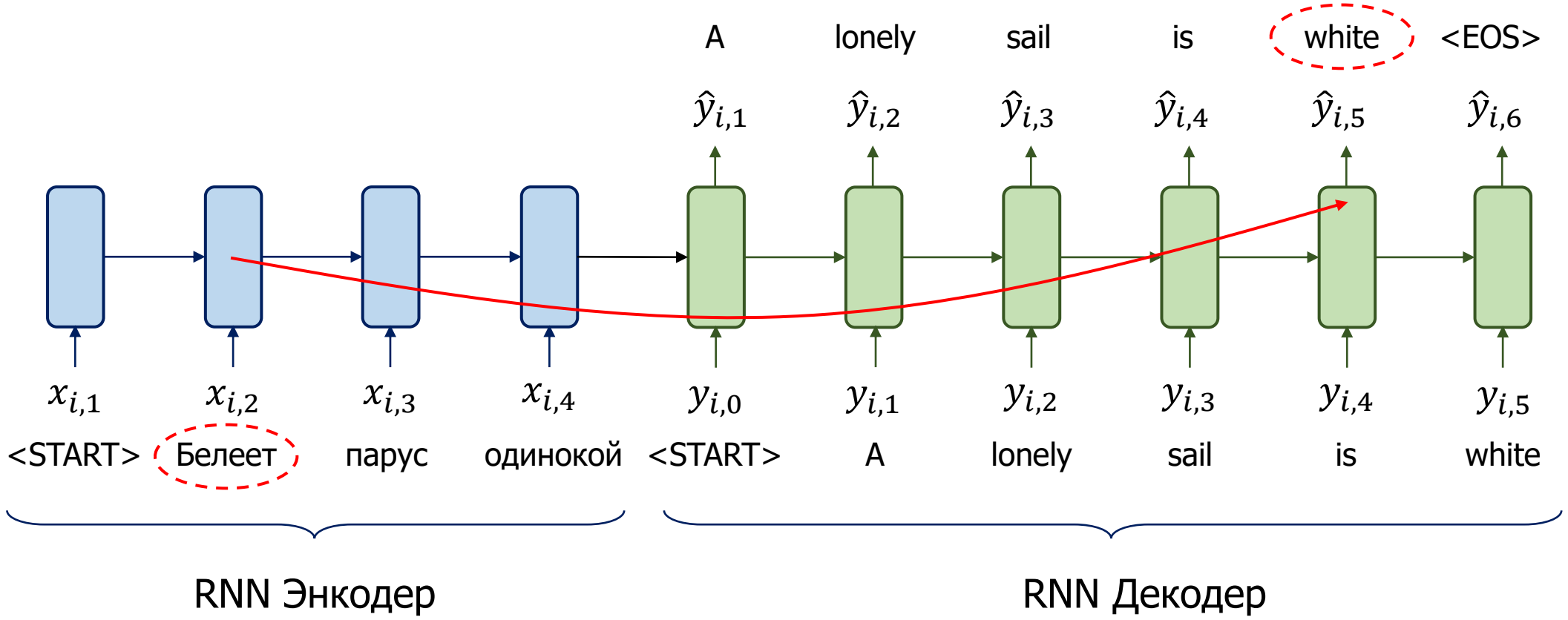
Модели Seq2Seq



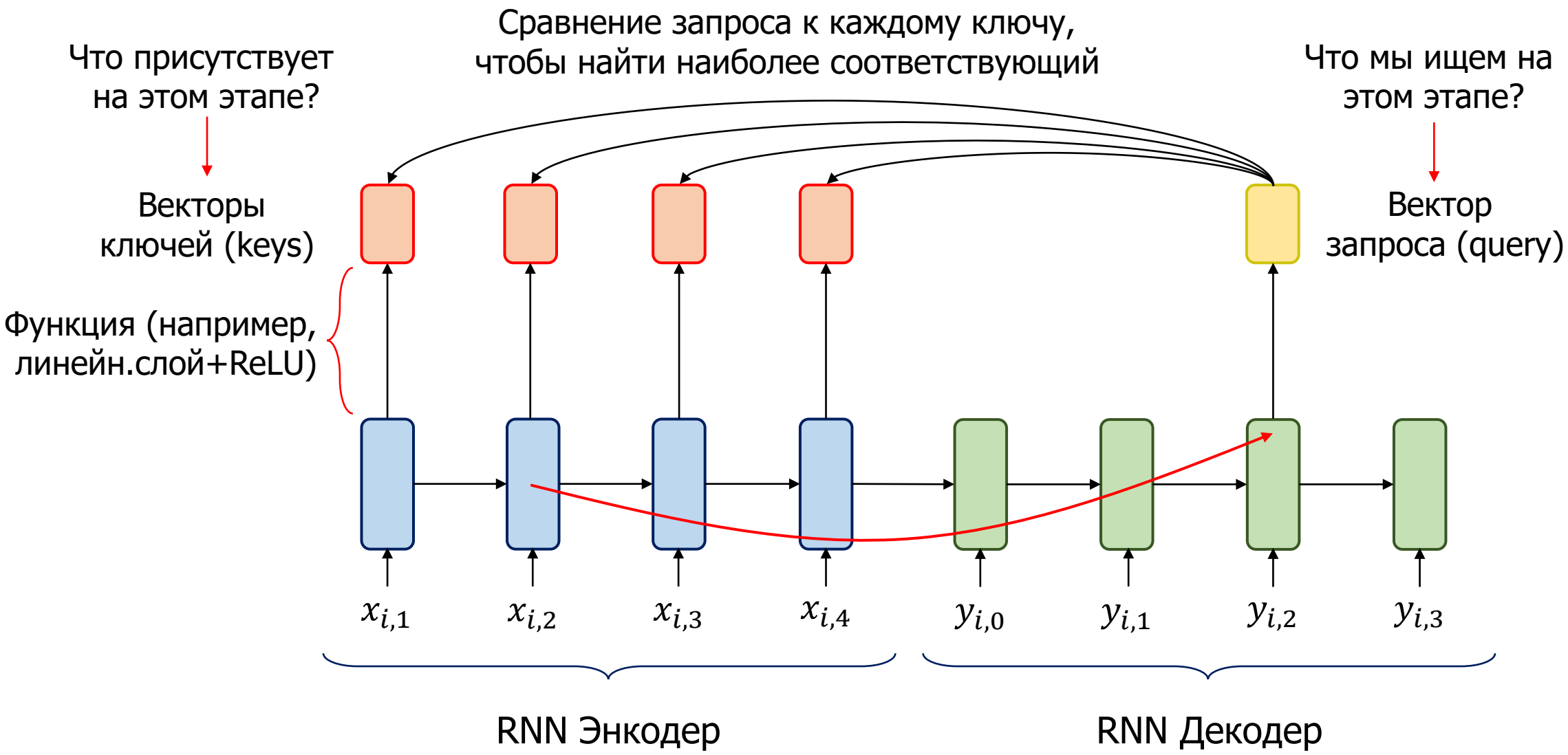
Использование стека рекуррентных слоёв



Проблемы при передаче контекста

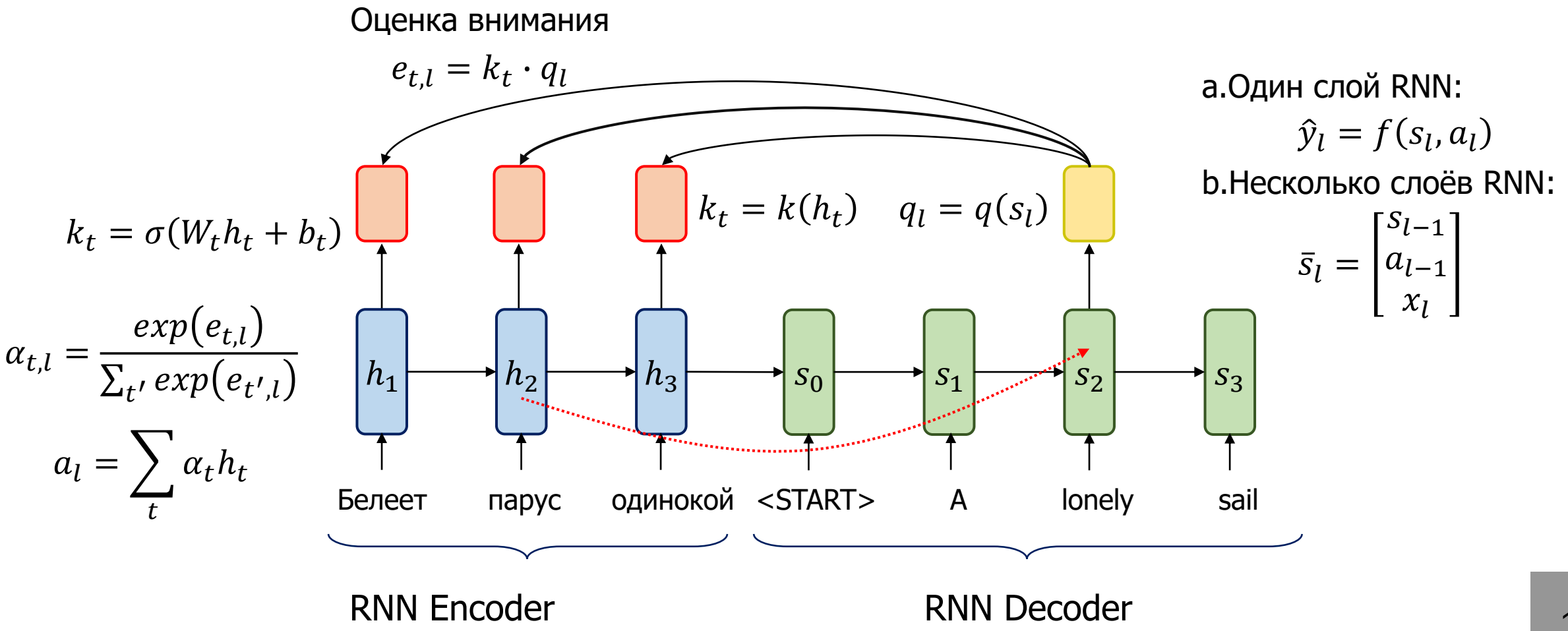


Ключи (keys) и запросы (queries)

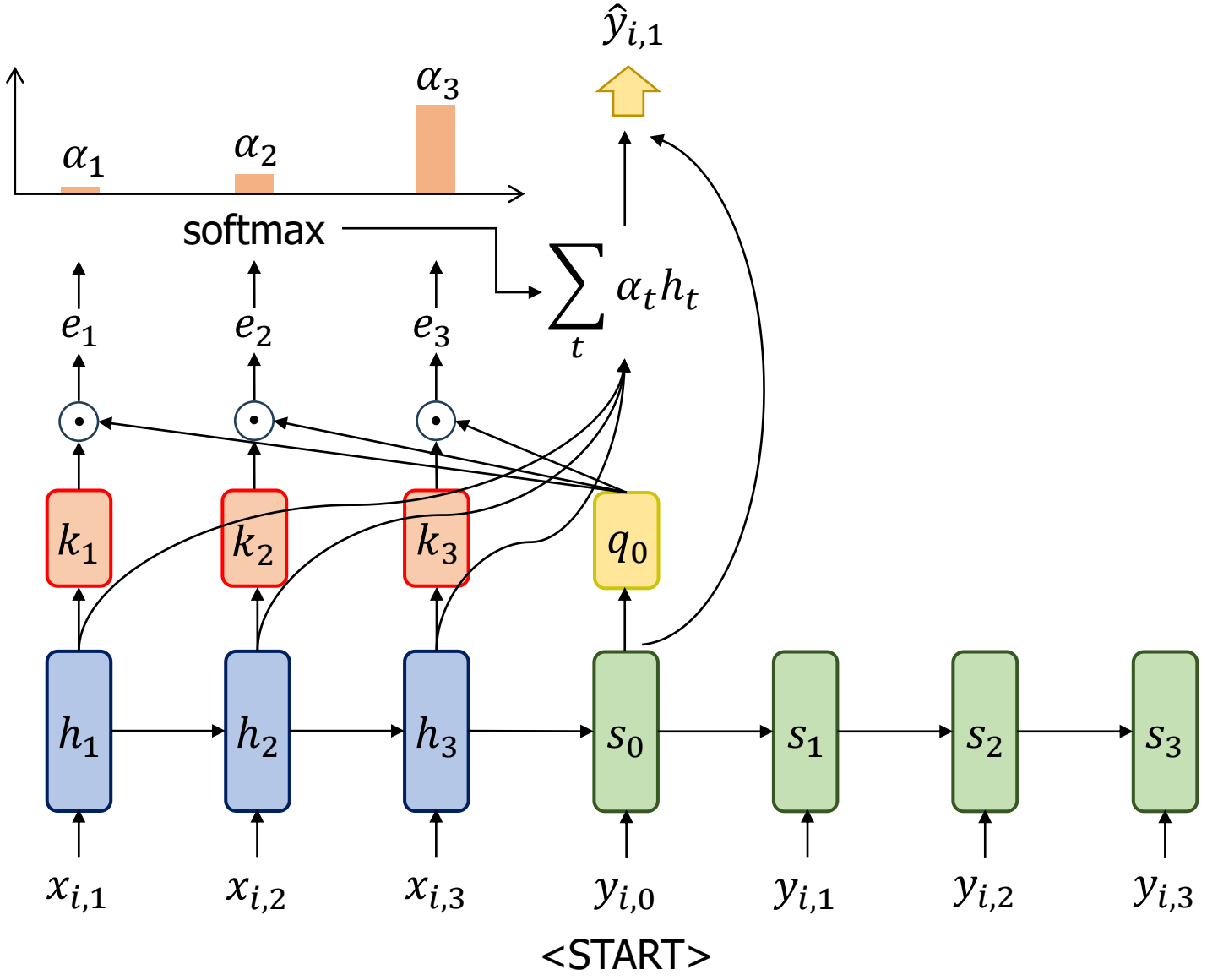


Механизм внимания (Attention)

Внимание присваивает важность каждому слову, вычисляя «мягкие» веса для числового представления слова в контекстном окне.



Внимание в деталях



Варианты внимания

1. Простой вариант ключа и запроса

$$k_t = h_t \qquad q_l = s_l \qquad e_{t,l} = k_t \cdot q_l$$

$$\alpha_{t,l} = \frac{exp(e_{t,l})}{\sum_{t'} exp(e_{t',l})}$$

2. Линейное мультипликативное внимание

$$k_t = W_k h_t \quad q_l = W_q s_l \qquad e_{t,l} = h_t^T W_k^T W_q s_l = h_t^T W_e s_l$$

$$a_l = \sum_t \alpha_t h_t$$

3. Использование дополнительной функции

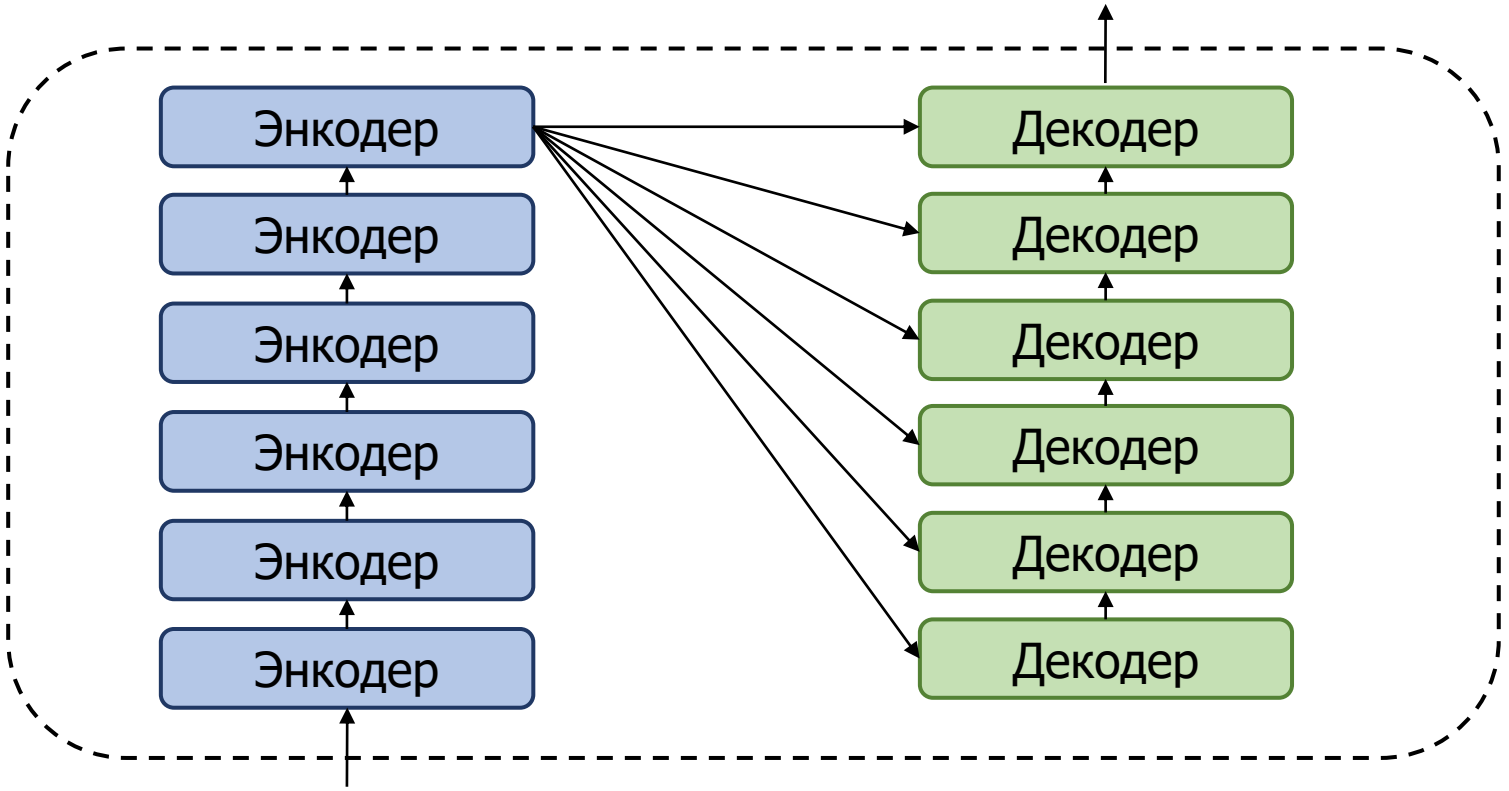
$$k_t = k(h_t) \quad q_l = q(s_l) \qquad e_{t,l} = k_t \cdot q_l$$

$$\alpha_{t,l} = \frac{exp(e_{t,l})}{\sum_{t'} exp(e_{t',l})} \qquad a_l = \sum_t \alpha_t v(h_t)$$

Обученная функция значений

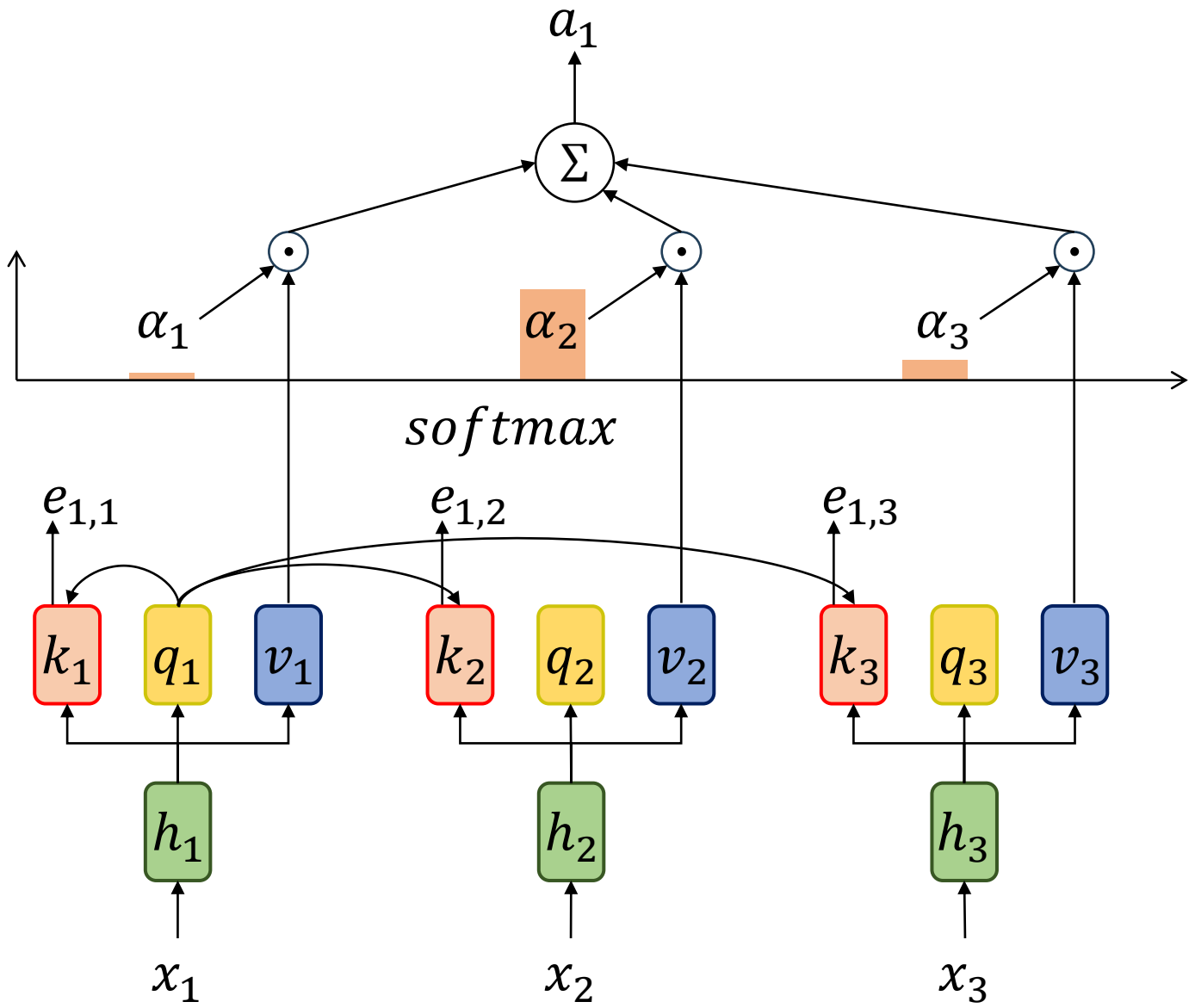
Общая структура трансформера

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



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

Самовнимание (Self-Attention)



$$a_l = \sum_t \alpha_{t,l} v_t = W_v \sum_t \alpha_{t,l} h_t$$

$$\alpha_{t,l} = \frac{exp(e_{t,l})}{\sum_{t'} exp(e_{t',l})}$$

$$e_{t,l} = k_t \cdot q_l$$

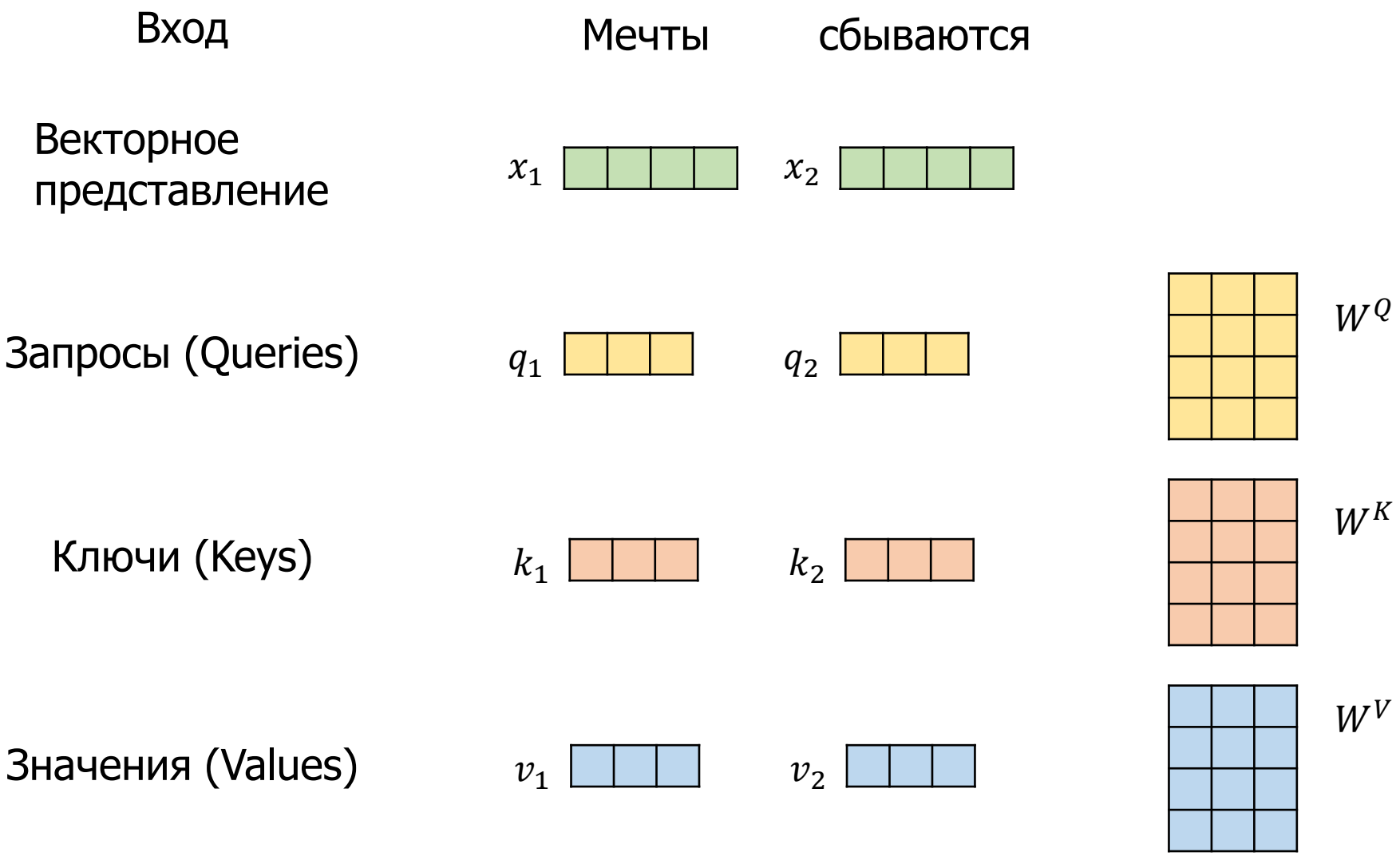
$$v_t = v(h_t) \qquad v_t = W_v h_t$$

$$k_t = k(h_t) \qquad k_t = W_k h_t$$

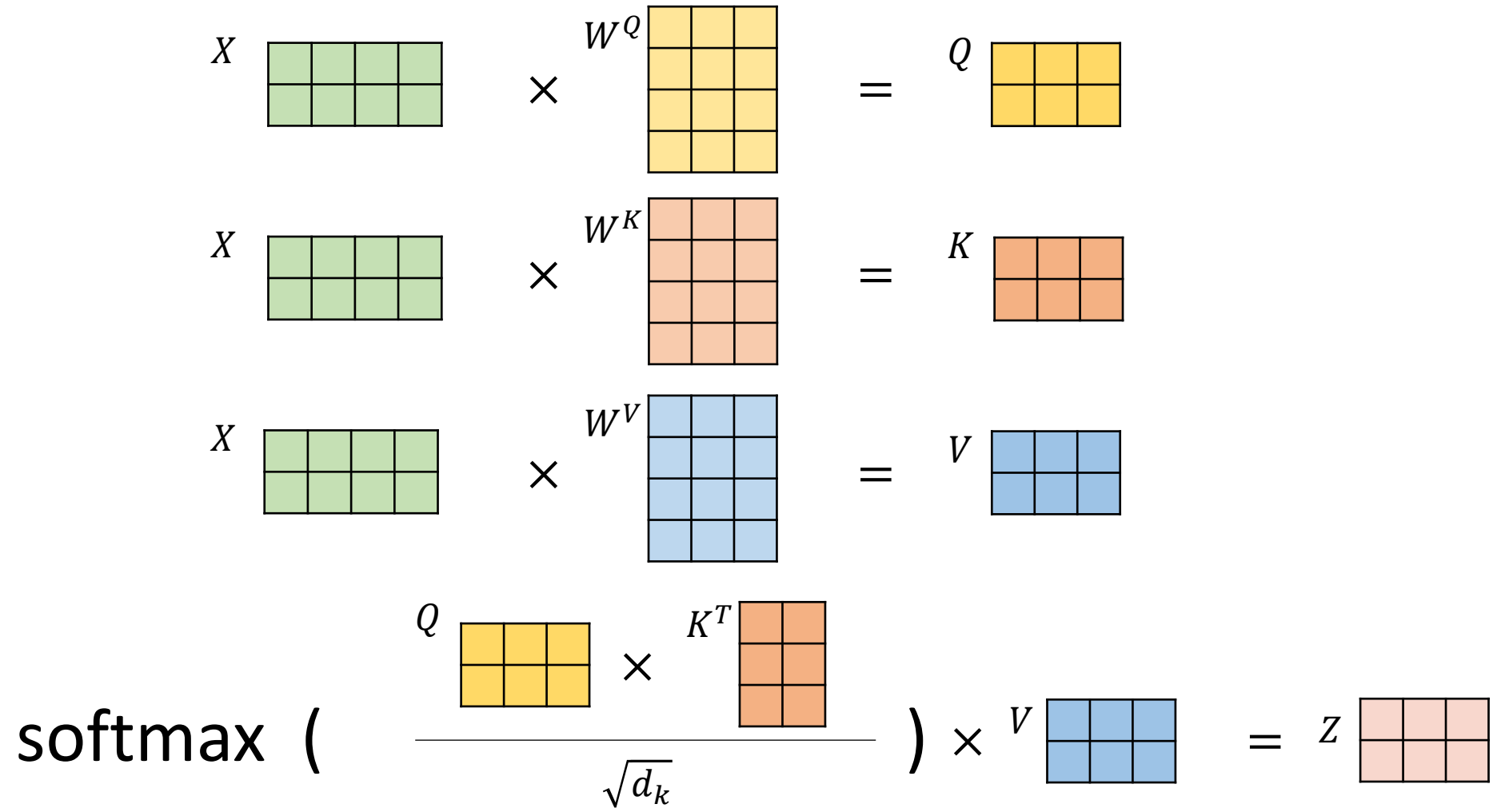
$$q_t = q(h_t) \qquad q_t = W_q h_t$$

$$h_t = \sigma(W x_t + b)$$

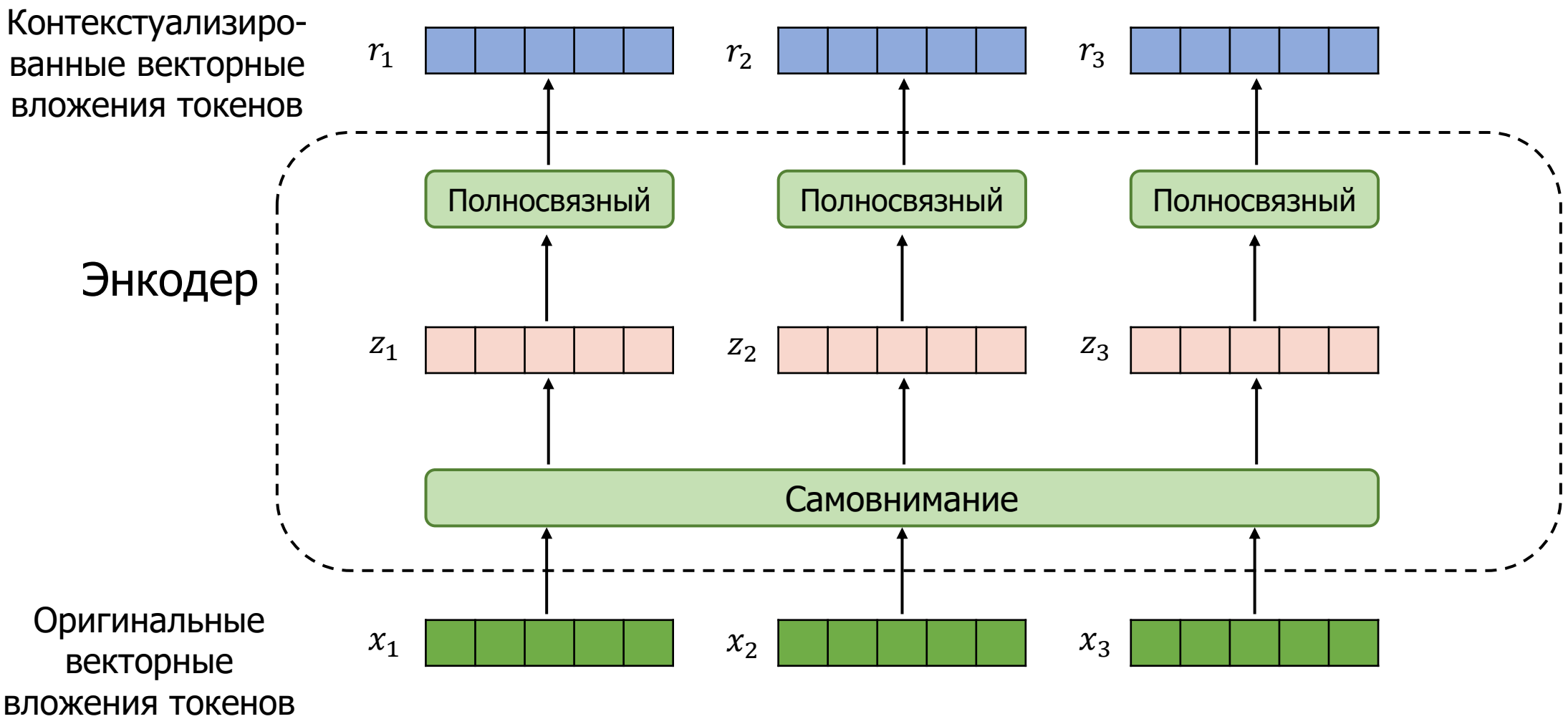
Вектора Query, Key и Value



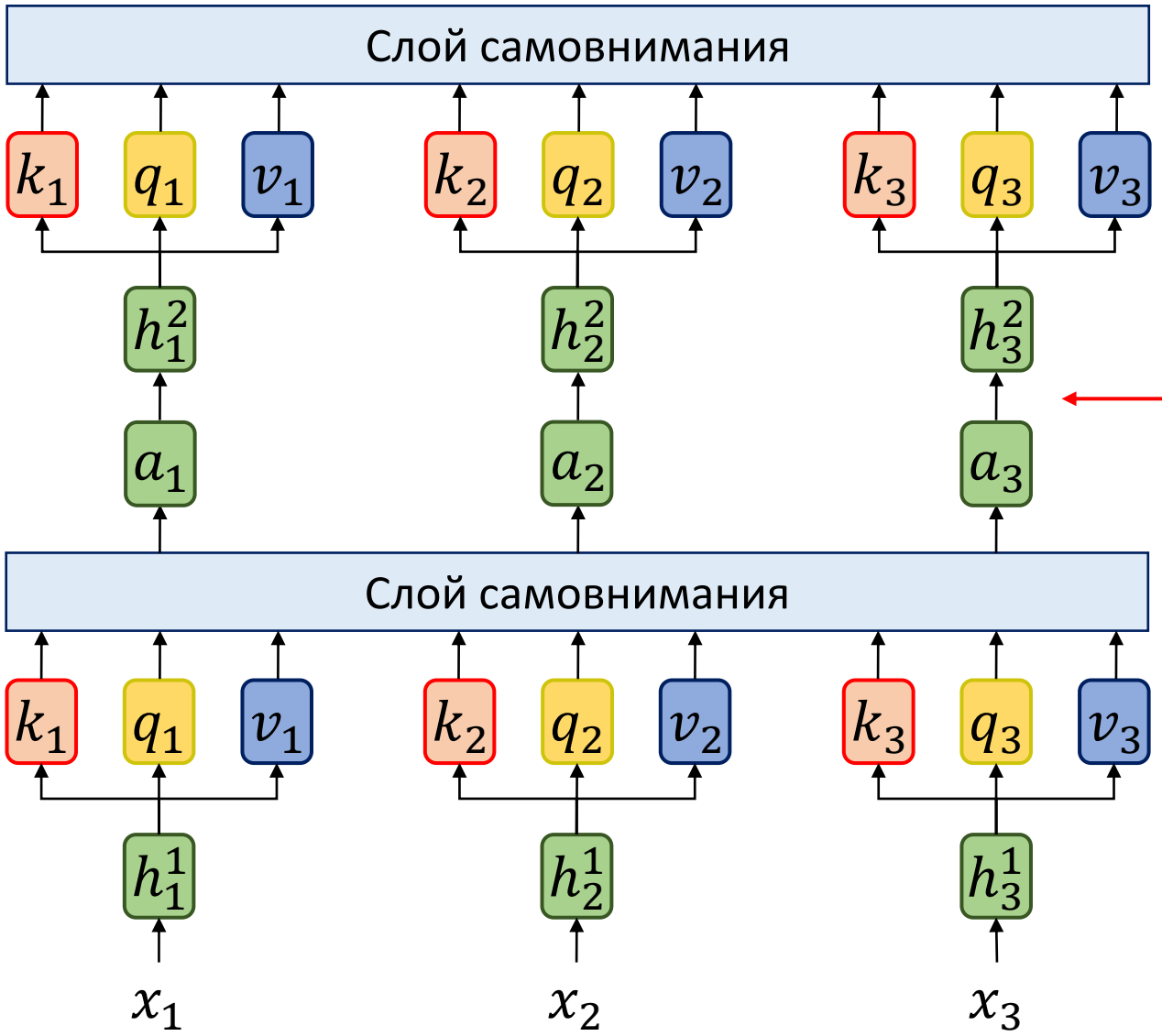
Матричные вычисления внутреннего внимания



Самовнимание в энкодере



Чередование внимания и нелинейности



Нелинейное преобразование

$$h_t^\ell = \sigma\left(W^\ell a_t^\ell + b^\ell\right)$$

Пример многоуровневого самовнимания

«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

Немного больше примеров

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.

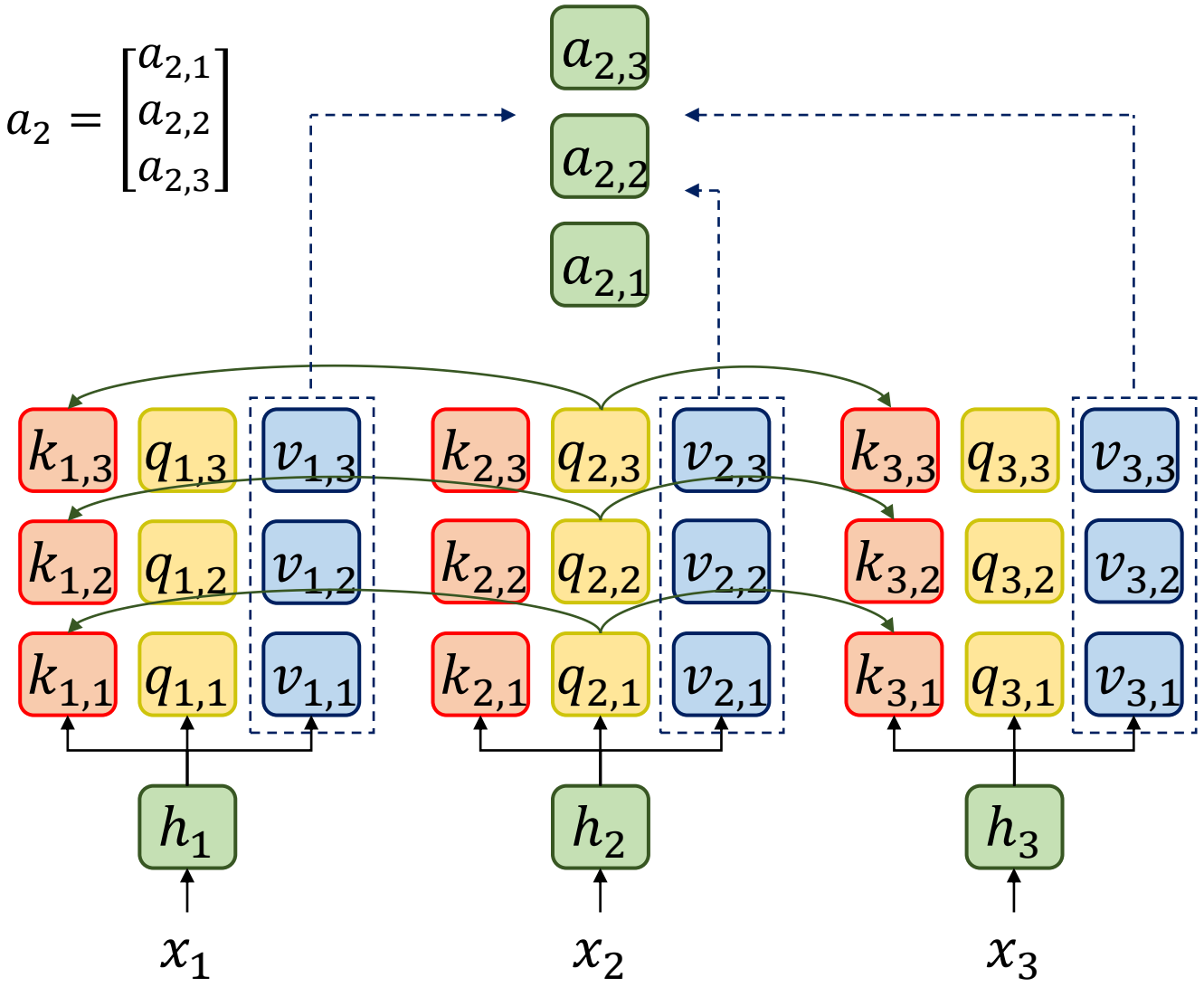
4.Sundays are always a little sad, but the day before is a sadder day.

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.

Многоголовое внимание



- ✓ Несколько векторов ключей, запросов и значения
- ✓ Расчет весов независимо для каждой головы внимания
- ✓ На практике 8 голов внимания показывают хороший результат

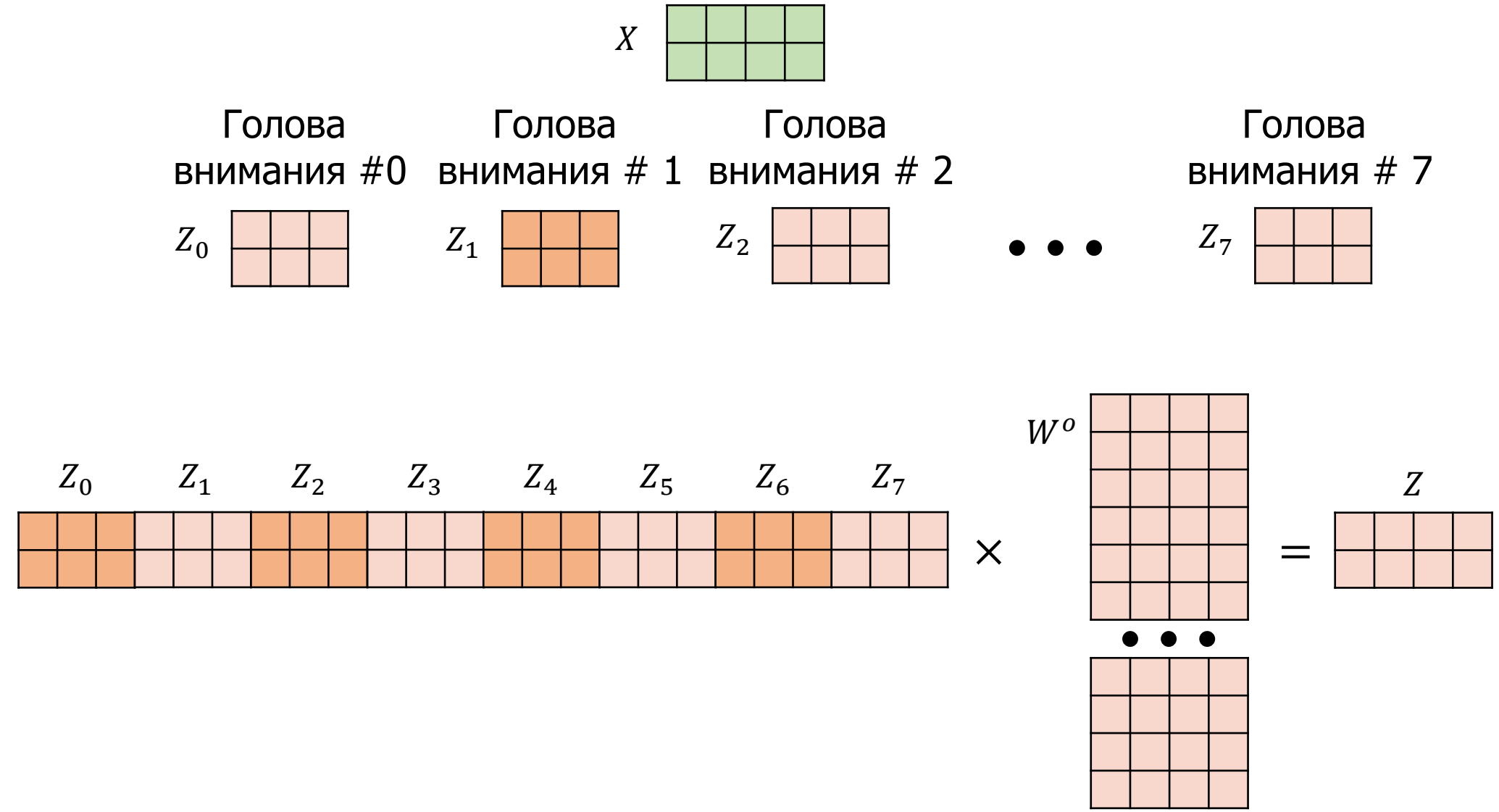
$$e_{t,l,i} = k_{t,i} \cdot q_{l,i}$$

$$\alpha_{t,l} = \frac{exp(e_{t,l,i})}{\sum_{t'} exp(e_{t',l,i})}$$

$$a_{l,i} = \sum_t \alpha_{t,l,i} v_{t,i}$$

Итоговый вектор внимания - конкатенация

Реализация множественного внимания



Позиционное кодирование

1.Наивное позиционное кодирование: $\bar{x}_t = \begin{bmatrix} x_t \\ t \end{bmatrix}$

2.Настроенное позиционное кодирование: $\bar{x}_t = \begin{bmatrix} x_t \\ p_t \end{bmatrix}$ или $\bar{x}_t = f(x_t, p_t)$

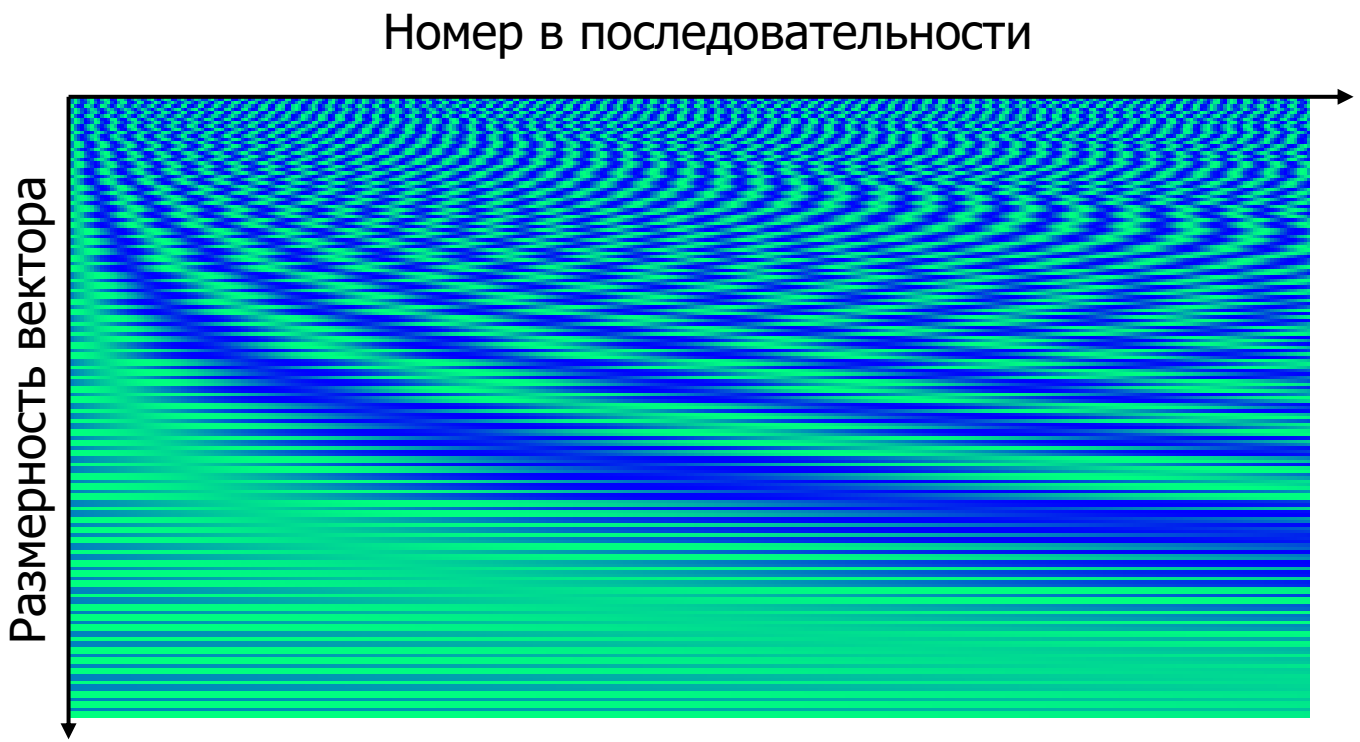
Относительная позиция токена более важна, чем его абсолютная позиция

«Моя свеча, бросая тусклый свет, в твой новый мир осветит бездорожье.» И.А. Бродский

«Ночь, улица, фонарь, аптека, Бессмысленный и тусклый свет.» А.А. Блок

Частотные представления

$$p_t = \begin{bmatrix} sin\left(t/1000^{2*1/d}\right) \\ cos\left(t/1000^{2*1/d}\right) \\ sin\left(t/1000^{2*2/d}\right) \\ cos\left(t/1000^{2*2/d}\right) \\ \dots \\ sin\left(t/1000^{2*\frac{d}{2}/d}\right) \\ cos\left(t/1000^{2*\frac{d}{2}/d}\right) \end{bmatrix}$$



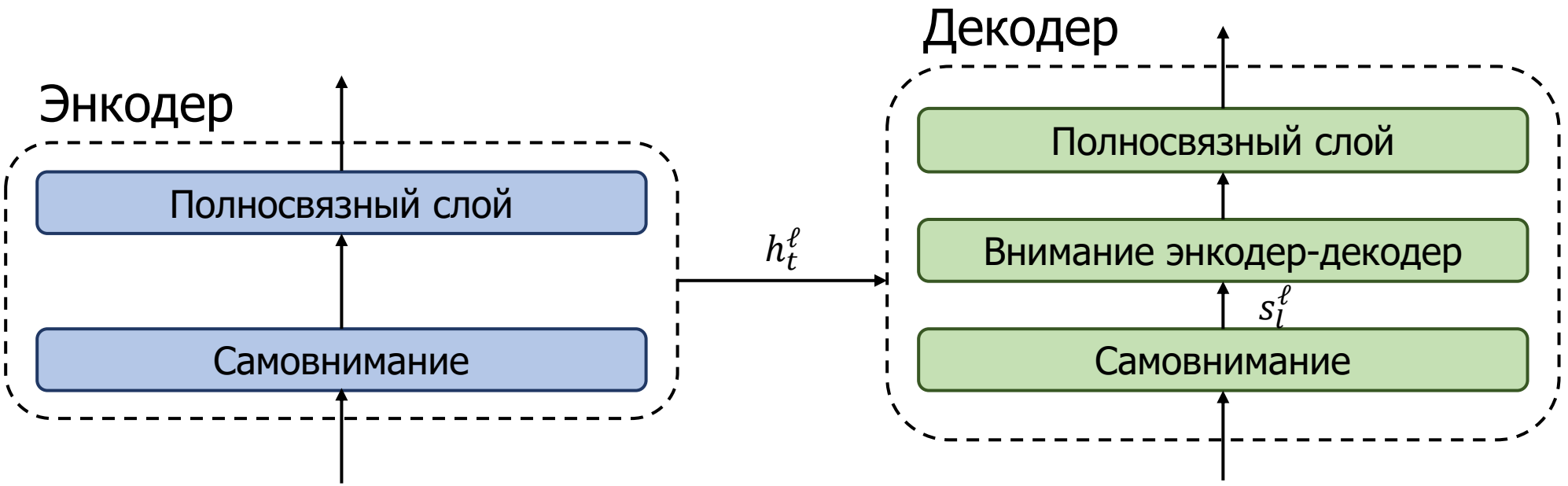
d - Размерность позиционного кодирования

Визуализация – векторных представлений позиционного кодирования

1.Конкатенация входа с позиционным кодированием: $\bar{x}_t = \begin{bmatrix} x_t \\ p_t \end{bmatrix}$

2.Сложение с эмбеддингом входа: $emb(x_t) + p_t$

Совместная работа энкодера и декодера



Запросы (query): $q_l^\ell = W_q^\ell s_l^\ell$

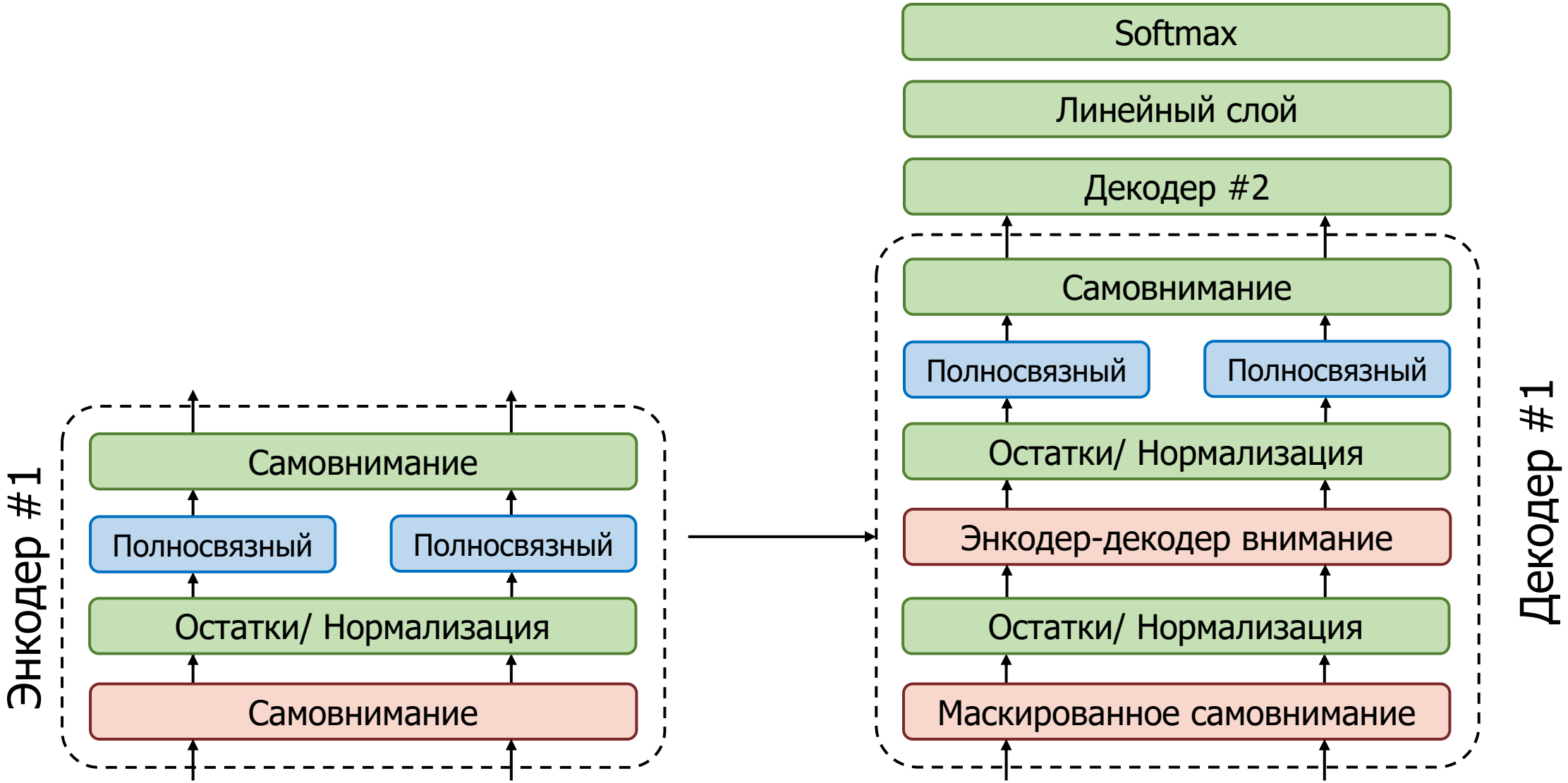
Ключи (key): $k_t^\ell = W_k^\ell h_t^\ell$

Значения (value): $v_t^\ell = W_k^\ell h_t^\ell$

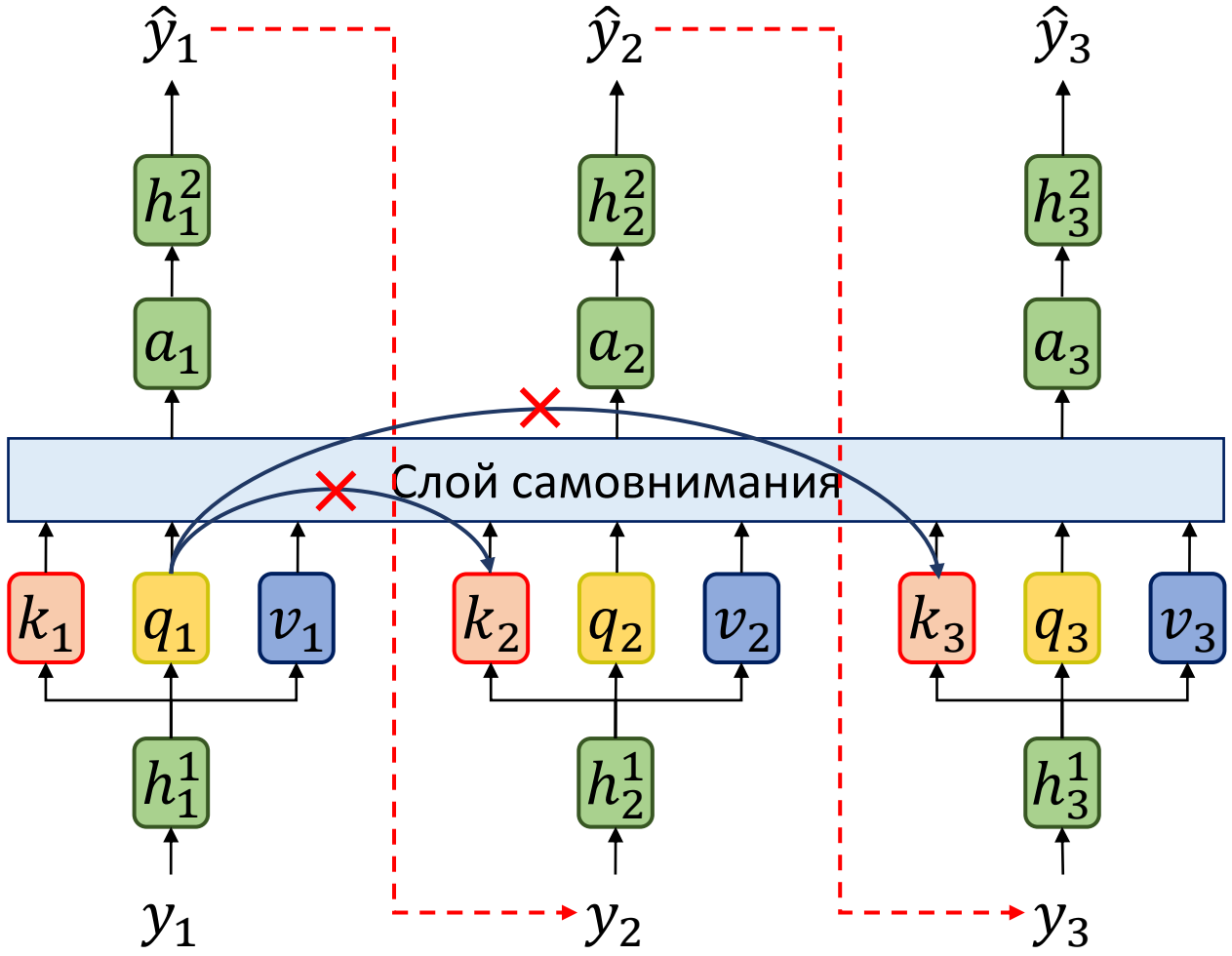
$$e_{l,t}^\ell = q_k^\ell k_t^\ell \qquad \alpha_{t,l}^\ell = \frac{exp\left(e_{t,l}^\ell\right)}{\sum_{t'} exp\left(e_{t',l}^\ell\right)}$$

Выход кросс-внимание: $c_l^\ell = \sum_t \alpha_{t,l}^\ell v_t^\ell$

Слои энкодера и декодера более детально



Маскированное внимание (прогноз следующего токена)



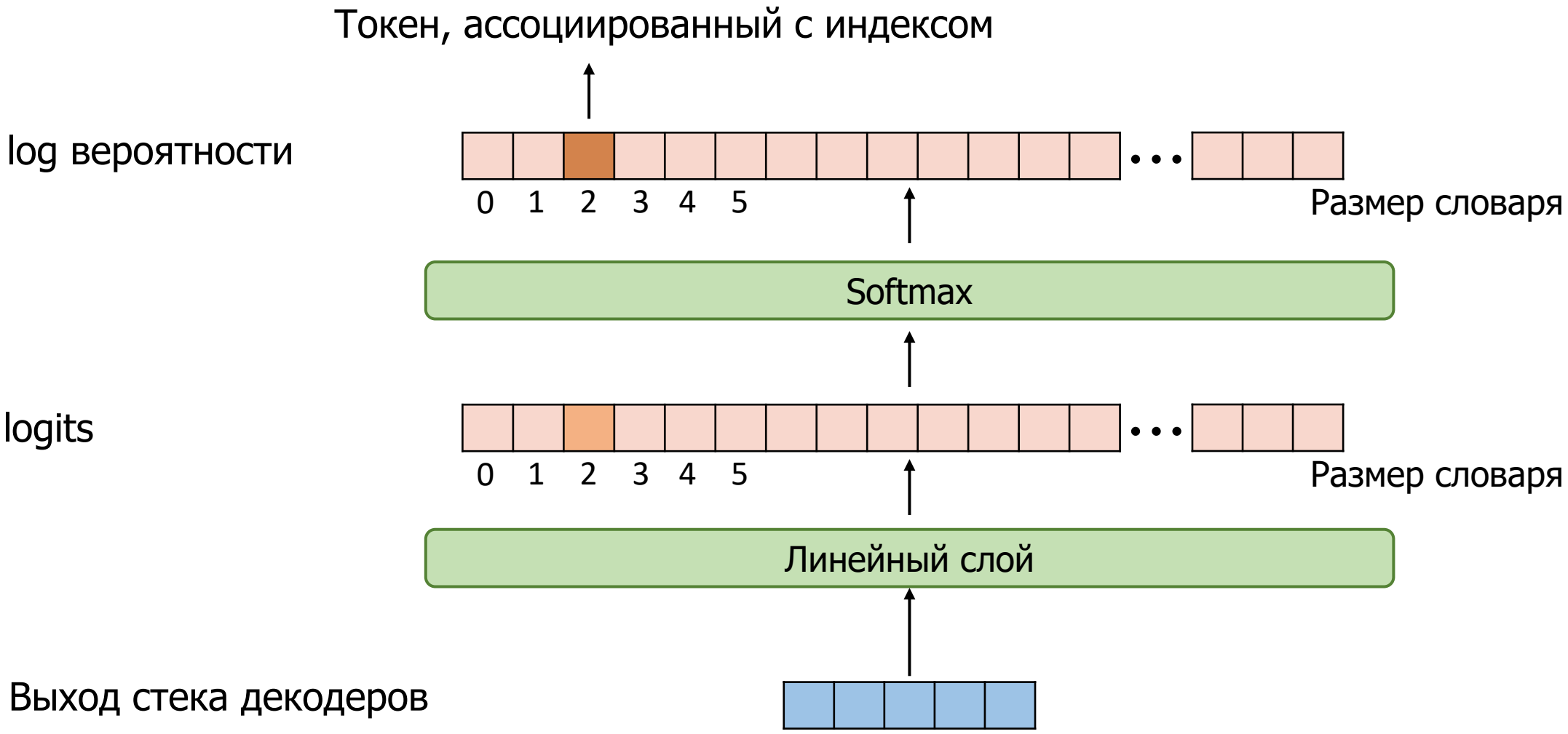
- ✓ Учет самовнимания на прошлых данных, но не на будущих (т.к. будущие ещё нужно сгенерировать)

$$e_{t,l} = \begin{cases} k_{t,i} \cdot q_{l,i}, l \geq t \\ -\infty, l < t \end{cases}$$

На практике:

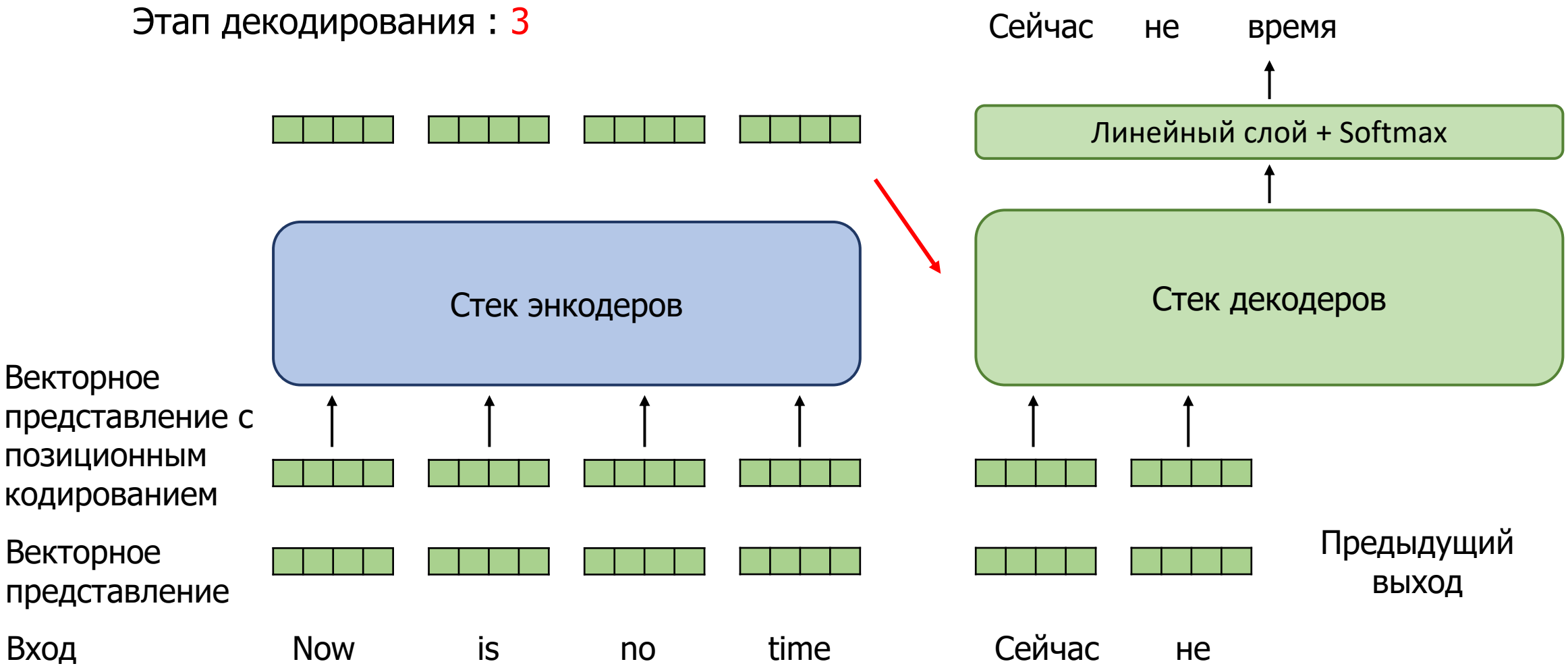
$$exp(e_{t,l}) \to 0, l < t$$

Финальные предсказания



Генерация последовательности

Этап декодирования : 3



"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)