



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

# ПРОГРАММНЫЙ ПРОЕКТ

## ВОПРОСНО-ОТВЕТНАЯ СИСТЕМА НА ОСНОВЕ НЕЙРОННОЙ СЕТИ

Выполнили студенты группы БПМИ-161  
Пономарев Вячеслав Андреевич  
Элбакян Мовсес Андраникович  
Ким Алёна Дмитриевна

Руководитель проекта:  
Симагин Денис Андреевич, ООО «Яндекс.Технологии», разработчик

Москва, 2019



# ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ

Разрабатывается вопросно-ответная система с пользовательским интерфейсом, доступная онлайн.

- Вопросно-ответная система – программный продукт, позволяющий пользователю найти ответ на свой вопрос, заданный в произвольной форме на естественном языке.



# АКТУАЛЬНОСТЬ РАБОТЫ

---

## Обоснование актуальности работы

- Поисковые системы в большинстве случаев предоставляют лишь выдачу релевантных страниц.
- Пользователь хочет сразу получить краткую информацию по своему вопросу, в идеале – точный ответ.
- Развитие области обработки естественного языка в задачах машинного обучения позволили проектировать качественные модели, способные решать задачи поиска ответа на вопрос в параграфе текста.

## Существующие программные решения

- The START Natural Language Question Answering System – MIT (1993)
- Answers - The Most Trusted Place for Answering Life's Questions (2005)
- Ask.com - What's Your Question? (1996)
- Семантическая поисковая система AskNet (2015)



# ЦЕЛЬ И ЗАДАЧИ РАБОТЫ

---

## Цель работы

- Разработка продукта, способного отвечать на вопросы, заданные человеком на английском языке на любую тему.

## Задачи

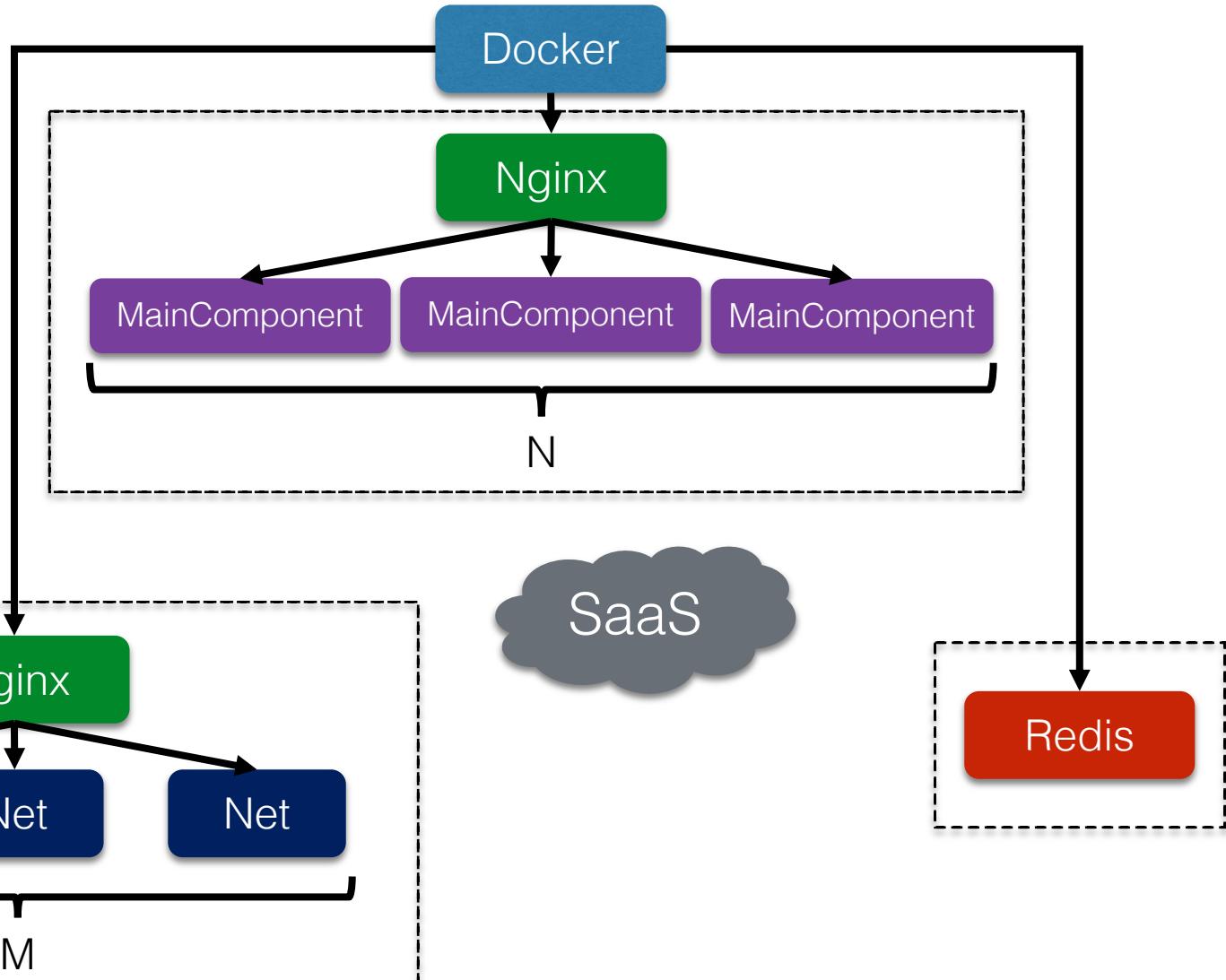
1. Разработать модуль предварительной обработки входных данных (Ким А.Д.)
2. Спроектировать и обучить модель машинного обучения, способную решать задачу поиска наиболее вероятного ответа на заданный вопрос в контексте (Пономарев В.А.)
3. Разработать и запустить серверную компоненту системы, обеспечить доступность системы онлайн (Элбакян М.А.)
4. Обеспечить возможность поиска наиболее релевантных документов в базе знаний при отсутствии предоставленного контекста (Элбакян М.А.)
5. Разработать удобный в использовании интерфейс системы (Ким А.Д.)

# АРХИТЕКТУРА СЕРВЕРНОЙ КОМПОНЕНТЫ

Элбакян Мовсес



# СХЕМА АРХИТЕКТУРЫ





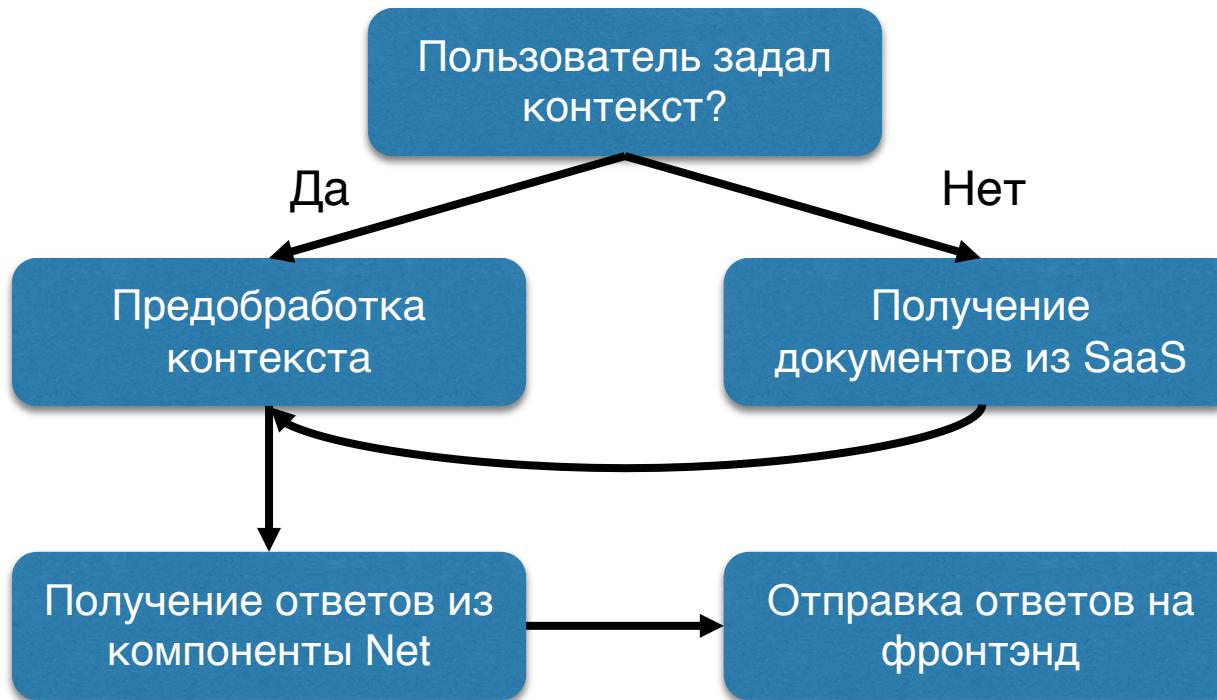
# NGINX



- Проксирует входной траффик
- Балансирует нагрузку на экземпляры компонент
- Служит точкой раздачи статических файлов фронтэнда



# MAINCOMPONENT. БИЗНЕС-ЛОГИКА





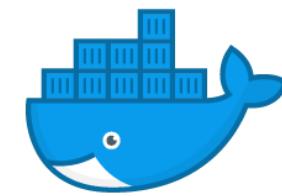
# MAINCOMPONENT. ТЕХНОЛОГИИ



aiohttp



aiowiki



docker

- Запущено N экземпляров компоненты с помощью supervisord
- aiohttp позволяет асинхронно выполнять запросы по сети
- aiowiki скачивает и парсит документы с Википедии асинхронно



# MAINCOMPONENT + REDIS



- Попадание в кэш сокращает время ответа в среднем на 2 секунды
- Отказоустойчивость за счет репликации
- Запуск происходит с помощью Docker. Поэтому легко горизонтально масштабируется
- В качестве Python-клиента используется библиотека aioredis



# NET

NGINX



- Запущено M экземпляров компоненты с помощью supervisord
- Flask-серверы очень легковесные
- Инференс на GPU



# SEARCH AS A SERVICE (SAAS)

---

The Google logo is displayed in its classic multi-colored font. The letters are arranged in a bold, sans-serif style. The colors used are blue for 'G', red for 'oo', yellow for 'g', green for 'l', and red for 'e'.

- Используется Google API
- Ответ в виде JSON
- Высококачественный поиск документов на английском

# МОДУЛЬ ОБРАБОТКИ ДАННЫХ

Ким Алёна



# ДАННЫЕ ДЛЯ ОБУЧЕНИЯ

Набор данных SQuAD 2.0 (*Rajpurkar et al., 2018*)

- Более 500 статей из Википедии
- Более 100 000 пар контекст-вопрос из SQuAD 1.1
- Дополнительно 50 000 пар контекст-вопрос без ответа



<https://rajpurkar.github.io/SQuAD-explorer/>



# ПРЕДОБРАБОТКА ДАННЫХ

---

Варианты представления входных данных:

- Bag of Words
- Tf-idf
- Векторные представления слов (контекстно независимые)
- Векторные представления слов (контекстно зависимые)
- Генерация признаков:
  1. части речи;
  2. именованные сущности;
  3. статистические признаки;
  4. tf-статистики



# ПРЕДОБРАБОТКА ДАННЫХ

---

spaCy – библиотека для расширенной обработки естественного языка:

- токенизация текста;
- получение начальной формы слова;
- разметка частей речи (POS);
- разметка именованных сущностей (NER)

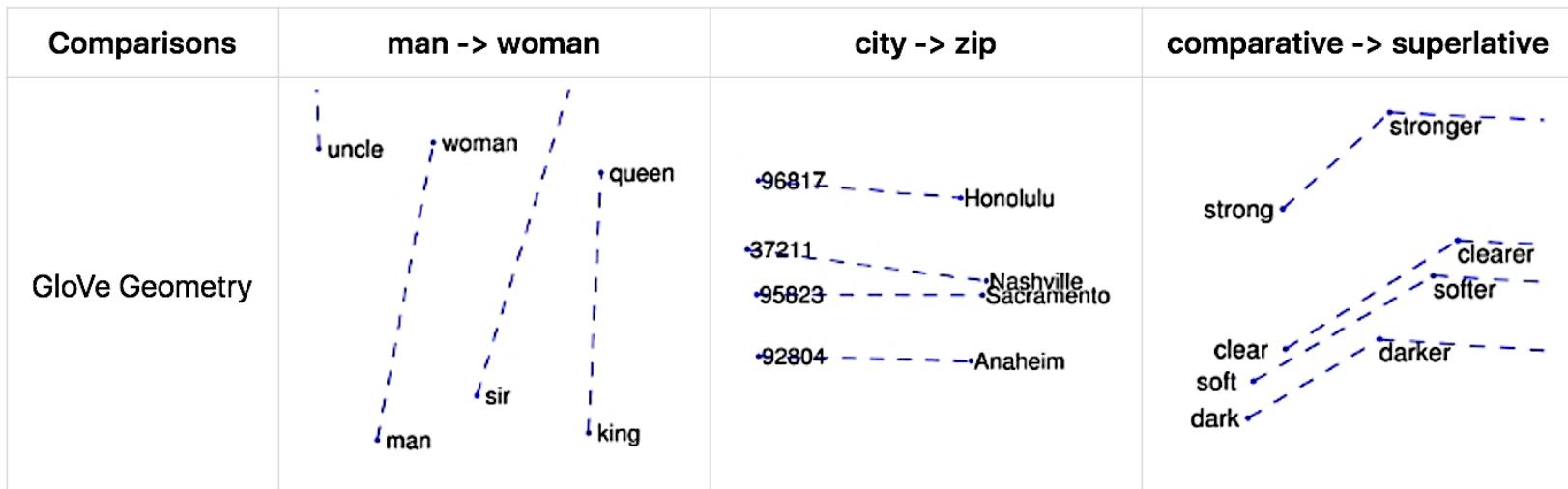
<https://spacy.io>



# GLOVE ПРЕДСТАВЛЕНИЯ

## GloVe: Global Vectors for Word Representations (Pennington et al., 2014)

nearest neighbors of <i>frog</i>	Litoria	Leptodactylidae	Rana	Eleutherodactylus
Pictures				

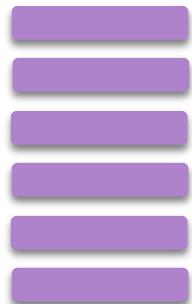


<https://nlp.stanford.edu/projects/glove>

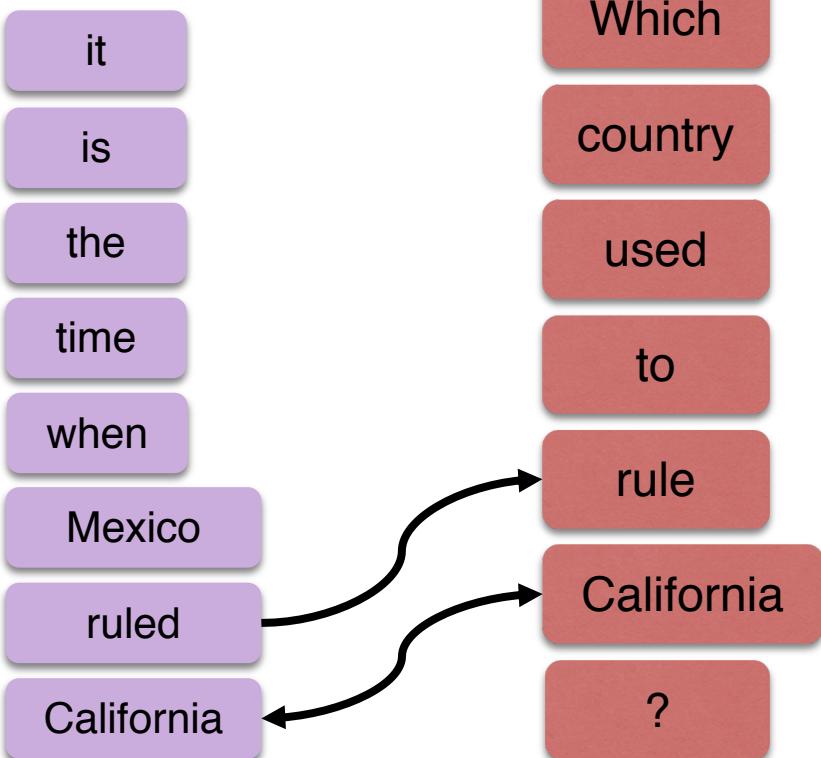
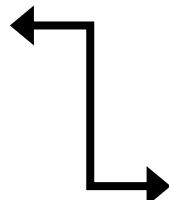
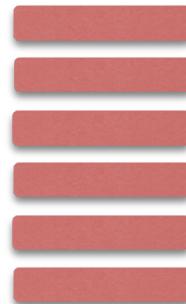


# ПЕРЕКРЕСТНЫЕ ПРИЗНАКИ

## Контекст



## Вопрос





# ПРЕДОБРАБОТКА ДАННЫХ

---

$$[GloVe_i, POS_i, NER_i, STATS_i]$$

- Обрабатывается каждая пара контекст-вопрос для обучающего и валидационных датасетов
- Реализована возможность частичной предобработки – считаются все признаки кроме перекрёстных

# АРХИТЕКТУРА НЕЙРОСЕТЕВОЙ МОДЕЛИ

Пономарев Вячеслав



# СУЩЕСТВУЮЩИЕ ПОДХОДЫ

- Обучение модели сразу под конкретную задачу
- Предобучение модели на большом объеме данных, дообучение под конкретную задачу

Одно из лучших решений – BERT (*Devlin et al., 2018*)

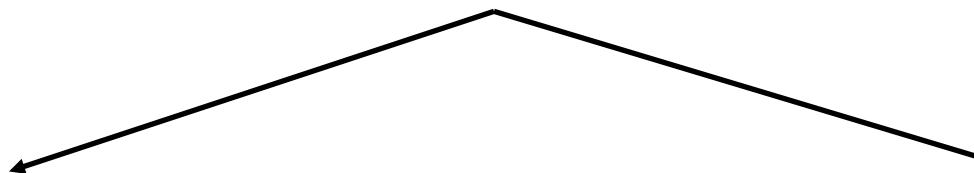
Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1	BERT + DAE + AoA (ensemble) <i>Joint Laboratory of HIT and iFLYTEK Research</i>	<b>87.147</b>	<b>89.474</b>

<https://rajpurkar.github.io/SQuAD-explorer/>



# СУЩЕСТВУЮЩИЕ ПОДХОДЫ

## Использование BERT



Дообучение готовой модели на датасете SQuAD

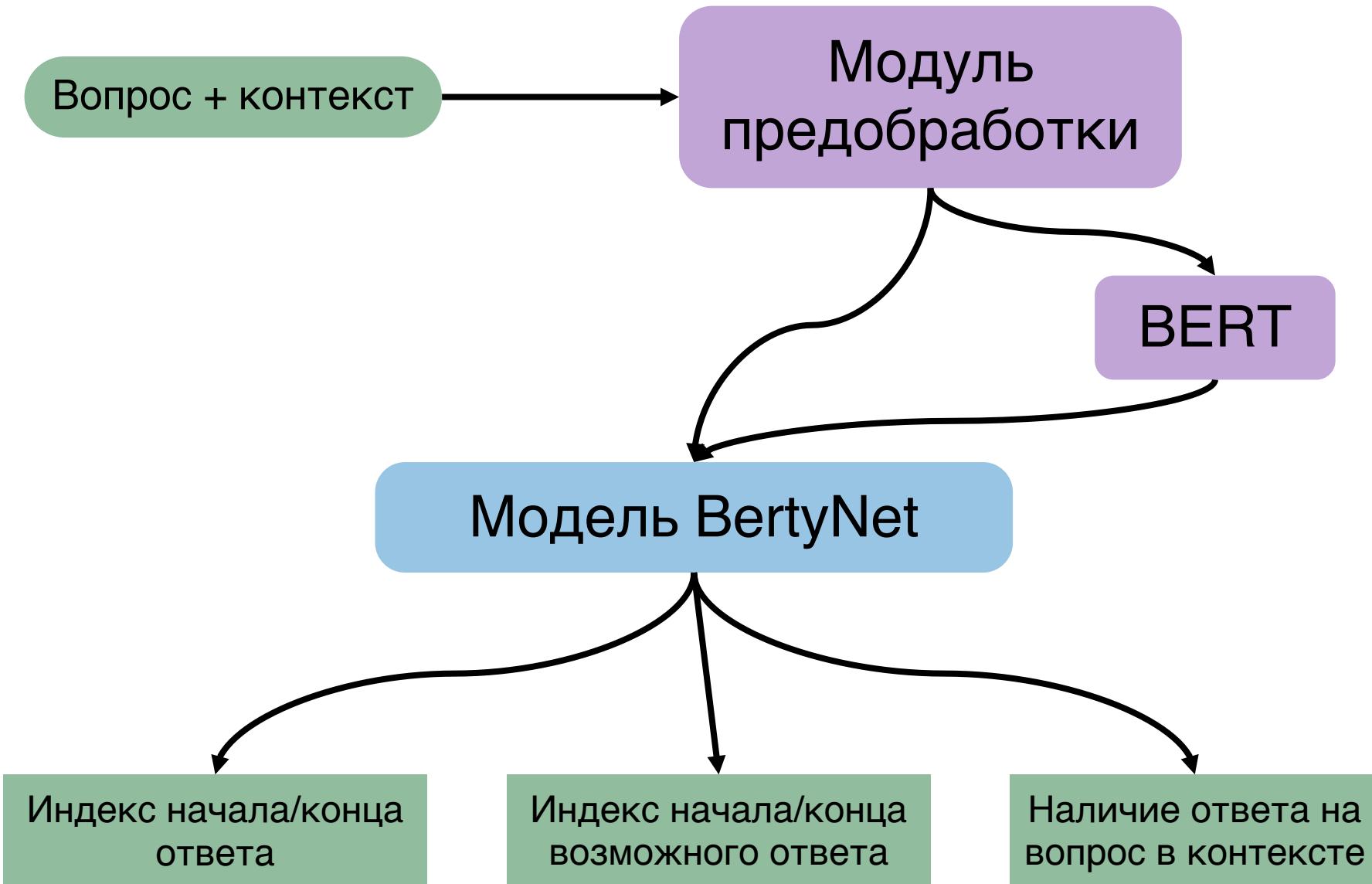
- + Возможно достичь наилучшего качества
  - Сильно вычислительно затратно
  - Невозможно привнести изменения в архитектуру

Извлечение векторных представлений слов и последующее их использование в другой модели

- + Вычислительно эффективно
  - Качество в основном зависит от модели, в которую подаются представления



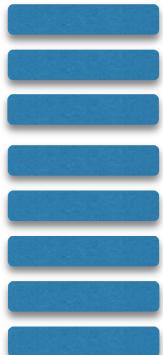
# ОБЩАЯ АРХИТЕКТУРА МОДЕЛИ





# ПОЛУЧЕНИЕ ПРЕДСТАВЛЕНИЙ BERT

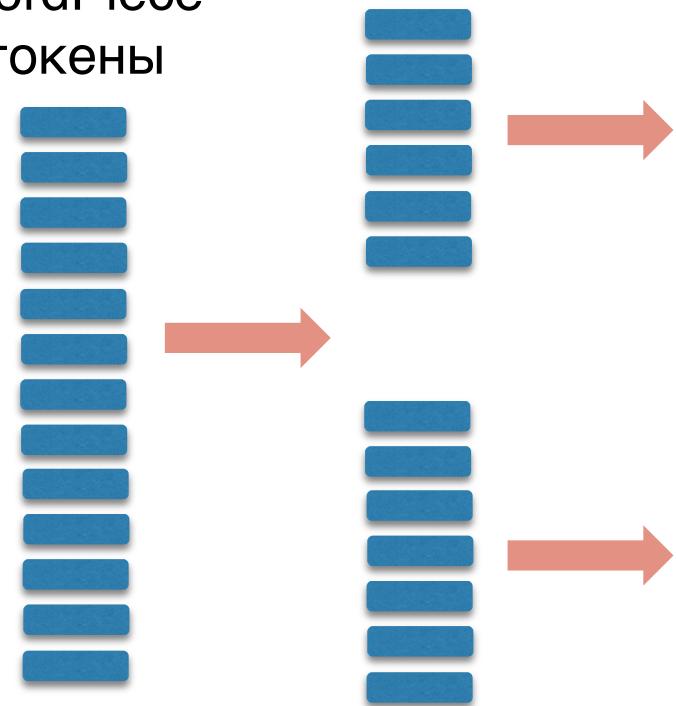
Исходные  
токены



WordPiece  
токенизатор



WordPiece  
токены



Отображение

[0, 2, 3, 4, 6, 8, 10, 12]

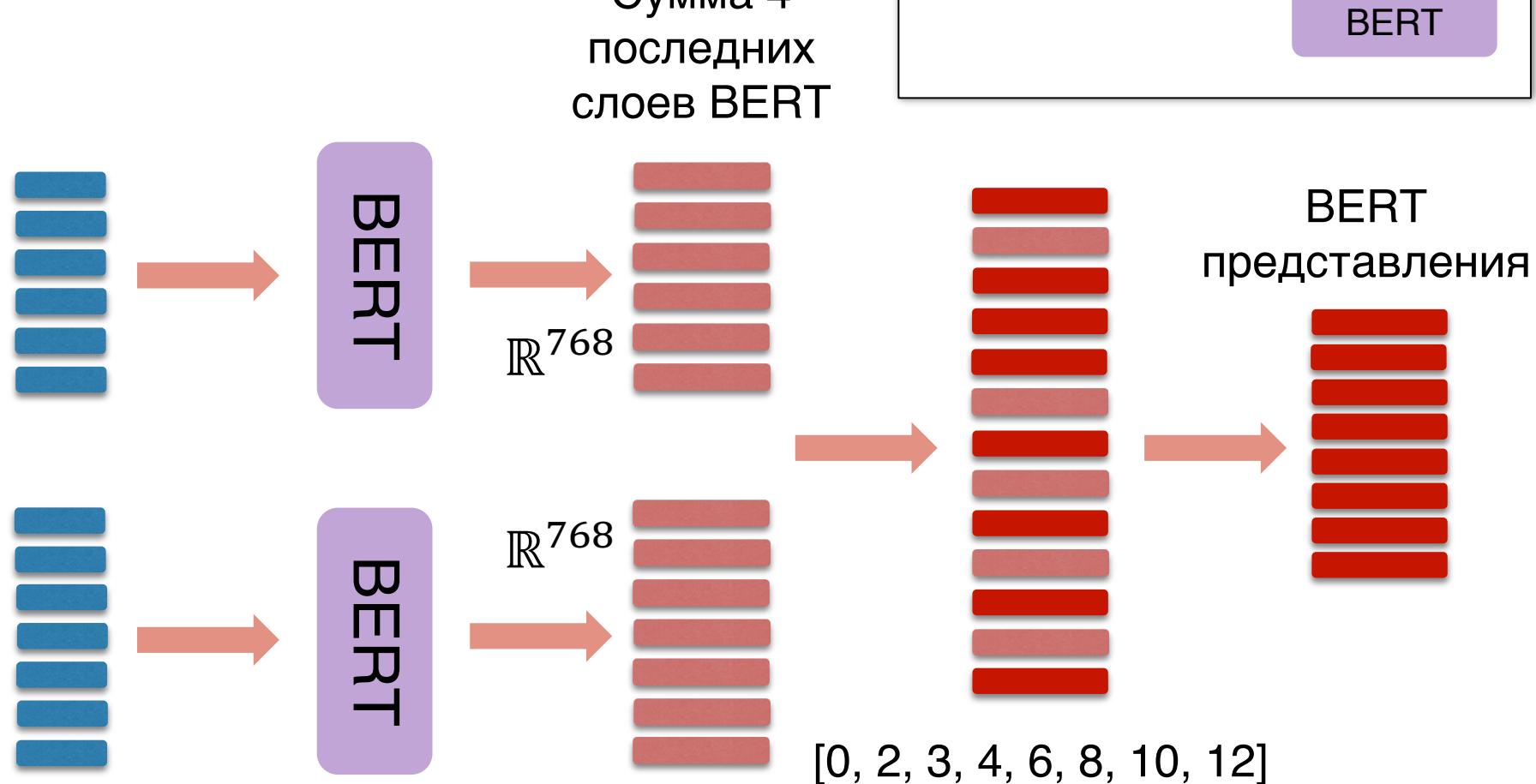
Модуль  
предобработки

BERT





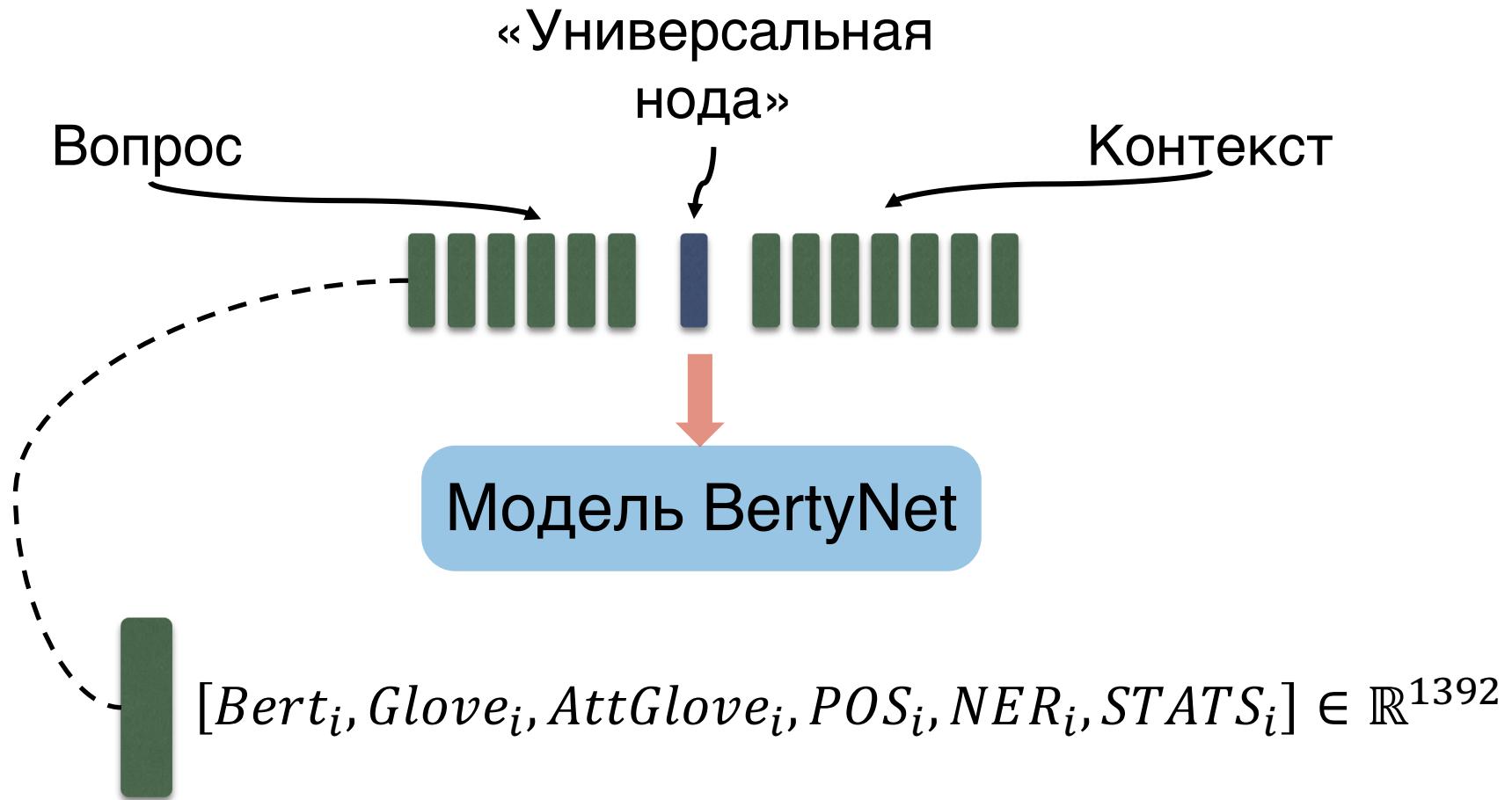
# ПОЛУЧЕНИЕ ПРЕДСТАВЛЕНИЙ BERT





# АРХИТЕКТУРА МОДЕЛИ

- Основа модели – U-Net (*Sun et al., 2018*)





# АРХИТЕКТУРА МОДЕЛИ

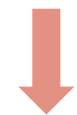


BiLSTM

BiLSTM

BiLSTM

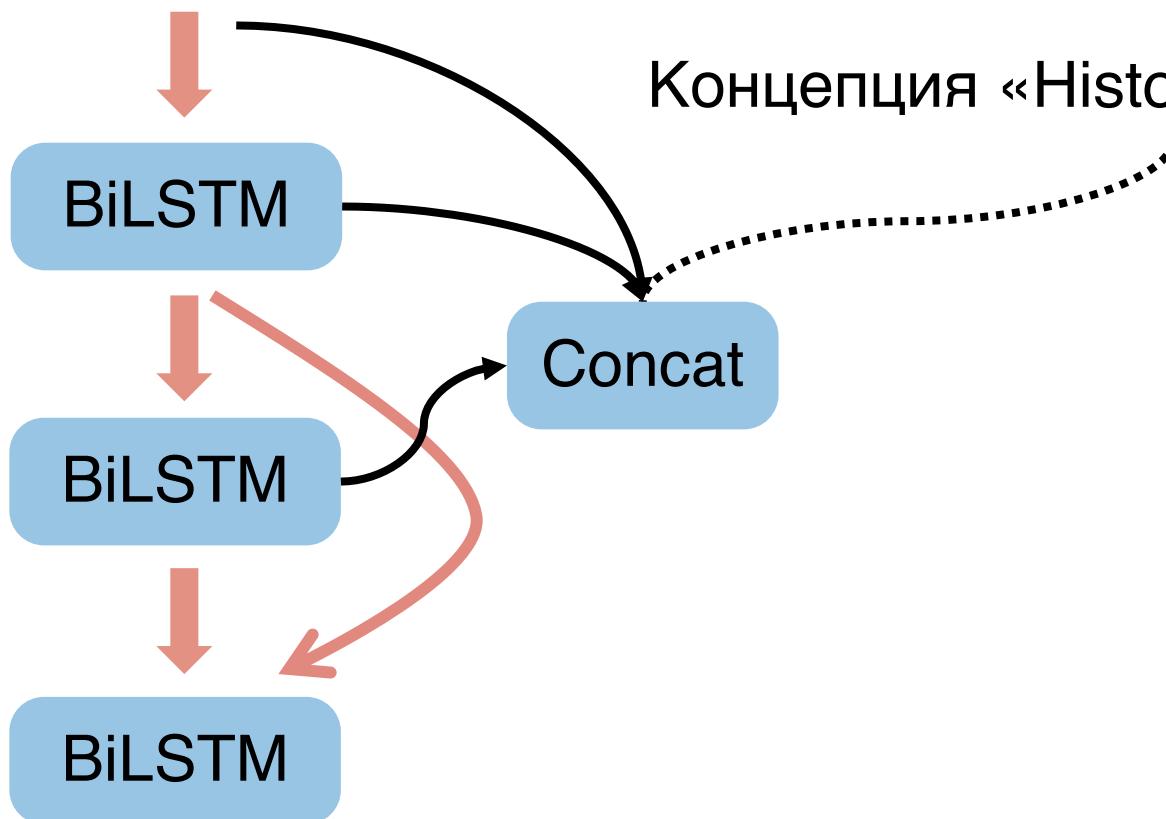
Модель BertyNet





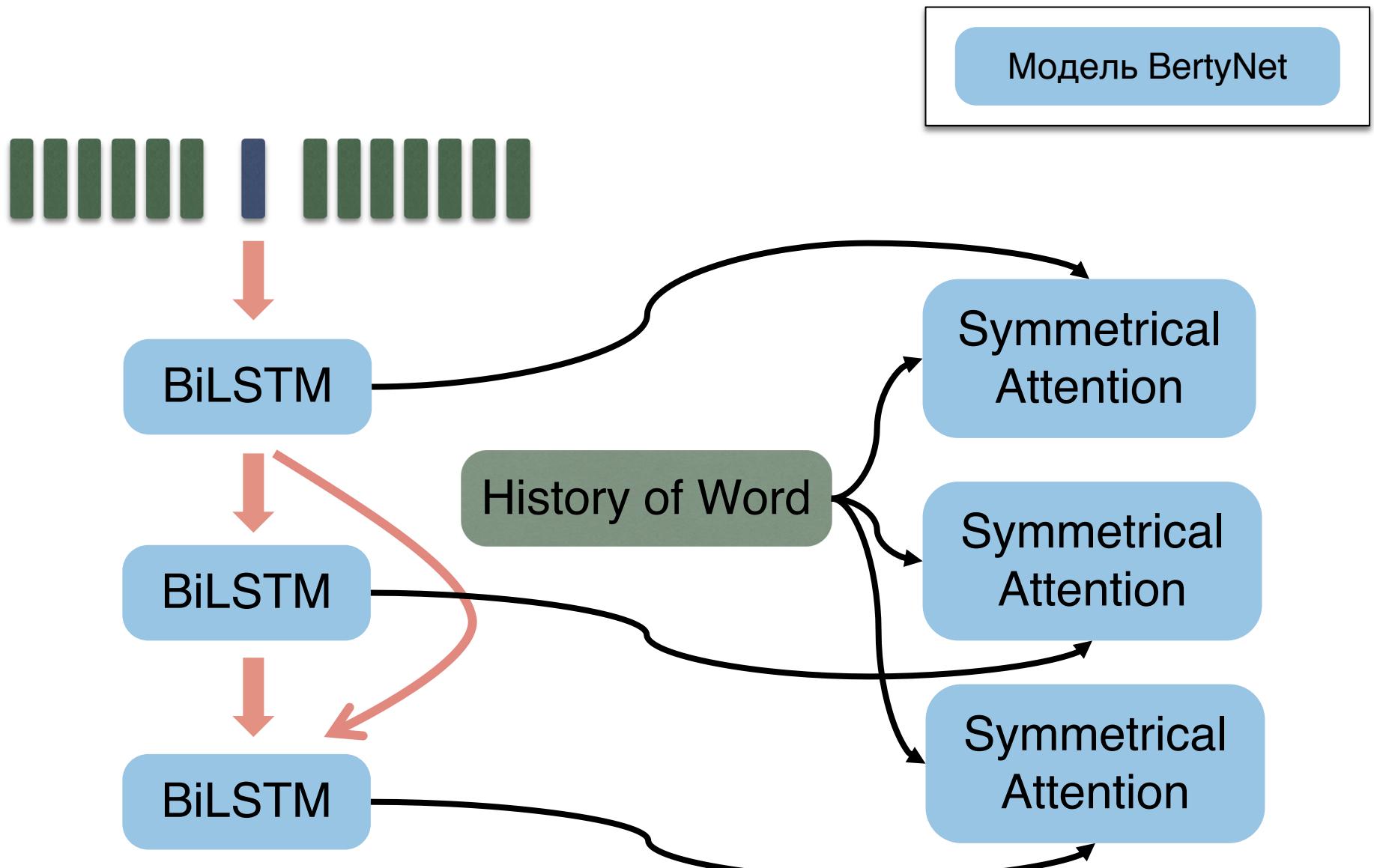
# АРХИТЕКТУРА МОДЕЛИ

Модель BertyNet





# АРХИТЕКТУРА МОДЕЛИ





# АРХИТЕКТУРА МОДЕЛИ

- Механизм внимания в симметричной форме с нелинейностью – FusionNet (*Huang et al., 2018*)

Symmetrical  
Attention

$$S = \left( \text{ReLU}(W \cdot \text{Ho}W_q) \right)^T \cdot D \cdot \text{ReLU}(W \cdot \text{Ho}W_p)$$

$$\widehat{H}_q = H_p \cdot \text{softmax}([S]^T)$$

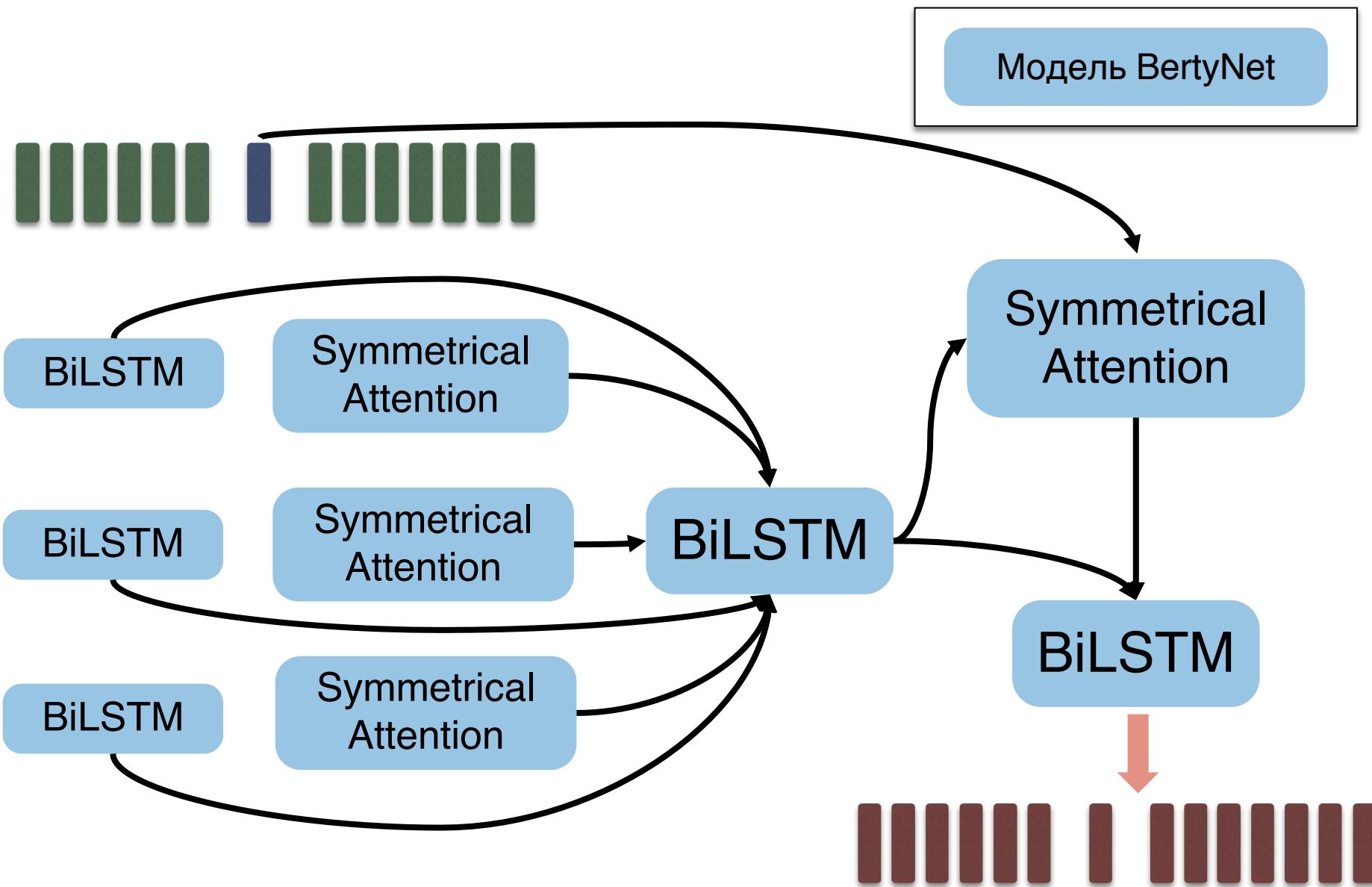
$$\widehat{H}_p = H_q \cdot \text{softmax}(S)$$

Где:

- $\text{Ho}W_p$  – матрица из векторов  $\text{Ho}W$ , соответствующих параграфу
- $\text{Ho}W_q$  – матрица из векторов  $\text{Ho}W$ , соответствующих вопросу
- $H_p$  – векторы состояний из LSTM слоя, соответствующие параграфу
- $H_q$  – векторы состояний из LSTM слоя, соответствующие вопросу
- $W$  – обучаемая матрица
- $D$  – обучаемая диагональная матрица

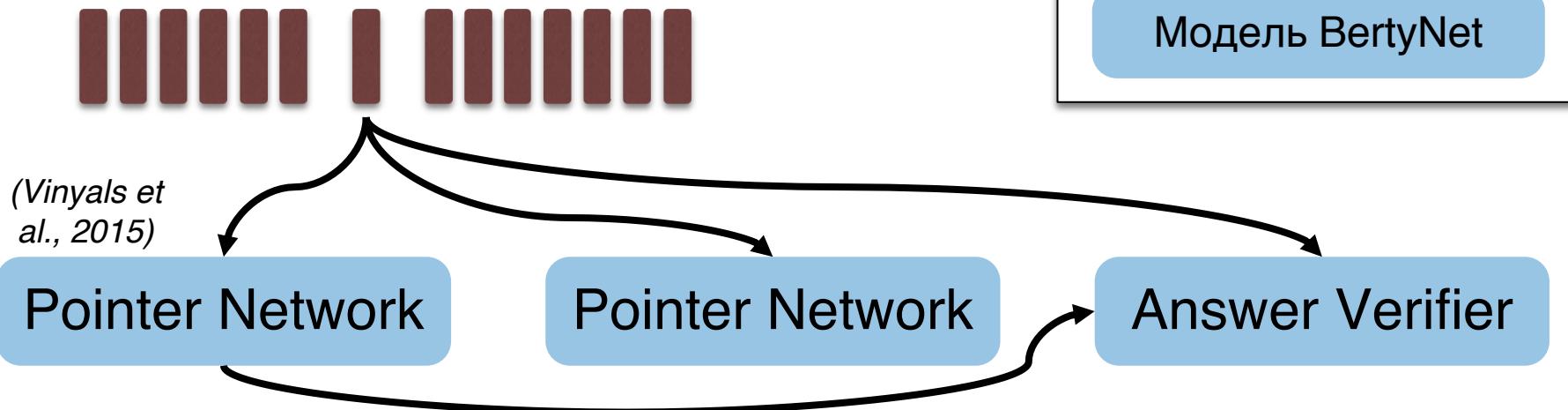


# АРХИТЕКТУРА МОДЕЛИ





# АРХИТЕКТУРА МОДЕЛИ



- Индекс токена – начала ответа
- Индекс токена – конца ответа
- Индекс токена – начала возможного ответа
- Индекс токена – конца возможного ответа
- Наличие ответа в контексте для заданного вопроса



# ОБУЧЕНИЕ МОДЕЛИ

Модель BertyNet

Pointer Network



$$L_A = -(\log s_a + \log e_b)$$

Pointer Network



$$L_{PA} = -(\log s_{plaus_{a^*}} + \log e_{plaus_{b^*}})$$

Answer Verifier



$$L_{AV} = -(\delta \log p_v + (1 - \delta) \log(1 - p_v))$$

- $s_i$  – оценка начала ответа для  $i$ -го токена
- $e_i$  – оценка конца ответа для  $i$ -го токена
- $a$  – верный индекс начала ответа
- $b$  – верный индекс конца ответа
- $p_v$  – оценка наличия ответа
- $\delta$  – существование ответа на самом деле

$$L_{total} = L_A + L_{PA} + L_{AV}$$



# РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

## Параметры обучения:

- Размер батча – 64
- Количество эпох – 30
- Вероятность dropout – 0.25
- Размер слоя механизма внимания (Symmetrical Attention) – 250
- Размер слоя BiLSTM – 125

## Параметры валидации:

- Нижняя граница уверенности в наличии ответа – 0.7
- Максимальное размер ответа – 15

Показатель	BertyNet (наша)	U-Net
F1 на dev выборке	70.4	<b>74.0</b>
F1 на dev выборке (только примеры с ответами)	<b>86.56</b>	<86



# ПРИМЕРЫ РАБОТЫ МОДЕЛИ

## Контекст

Alexander Sergeyevich Pushkin was a Russian poet, playwright, and novelist of the Romantic era who is considered by many to be the greatest Russian poet and the founder of modern Russian literature. <...> His novel in verse, Eugene Onegin, was serialized between 1825 and 1832. Pushkin **was fatally wounded in a duel** with his brother-in-law, Georges-Charles de Heeckeren d'Anthès, also known as Dantes-Gekkern, a French officer serving with the Chevalier Guard Regiment, who attempted to seduce the poet's wife, Natalia Pushkina.

## Вопрос

How did Pushkin die?

Ответ: wounded

Возможный ответ: fatally wounded in a duel

Уверенность в наличии ответа: 0.7015



# ПРИМЕРЫ РАБОТЫ МОДЕЛИ

## Контекст

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications...

## Вопрос

Why do people need machine learning?

Ответ: to make predictions or decisions...

Возможный ответ: is the scientific study of algorithms...

Уверенность в наличии ответа: 0.8941



# ПРИМЕРЫ РАБОТЫ МОДЕЛИ

## Контекст

The Moon is an astronomical body that orbits planet Earth and is Earth's only permanent natural satellite. <...> The Moon is thought to have formed about **4.51 billion years ago**, not long after Earth. <...>

The Moon's average orbital distance is **384,402 km** (238,856 mi), or 1.28 light-seconds. This is about **thirty** times the diameter of Earth. The Moon's apparent size in the sky is almost the same as that of the Sun, since the star is about 400 times the lunar distance and diameter...

## Вопрос

How far is the Moon from Earth?

Ответ: **4.51 billion years ago**

Возможный ответ: **thirty**

Уверенность в наличии ответа: **0.1068**

# ДЕМОНСТРАЦИЯ ИНТЕРФЕЙСА

Ким Алёна

[seann.ru](http://seann.ru)



# ВЫВОДЫ ПО РАБОТЕ

---

- Создан модуль, позволяющий обрабатывать тексты для дальнейшей работы с ними в нейросетевых моделях
- Проведены эксперименты по совмещению идей нескольких нейросетевых архитектур в одной
- Обучена модель, показывающее качество сравнимое с аналогичными архитектурами
- Спроектирован и реализован облачный сервис, позволяющий протестировать систему онлайн
- Реализована возможность автоматического поиска наиболее релевантного контекста по базе знаний
- Спроектирован удобный и информативный интерфейс для работы с системой



# НАПРАВЛЕНИЯ ДАЛЬНЕЙШЕЙ РАБОТЫ

---

- Модификация нейросетевой модели с целью достижения еще большего качества
- Увеличение производительности модели
- Улучшение производительности серверной компоненты
- Добавление опции выбора базы знаний
- Добавление опции отправления фидбека о выданных результатах



# СПИСОК ИСТОЧНИКОВ

1. Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [Электронный ресурс] // arXiv.org. URL: <https://arxiv.org/abs/1810.04805> (дата обращения: 9.02.19).
2. GloVe: Global Vectors for Word Representation [Электронный ресурс] // URL: <https://nlp.stanford.edu/projects/glove/> (дата обращения: 9.02.19).
3. Huang, H., Zhu, C., Shen, Y. and Chen, W. (2018). FusionNet: Fusing via Fully-Aware Attention with Application to Machine Comprehension. [Электронный ресурс] // arXiv.org. URL: <https://arxiv.org/abs/1711.07341> (дата обращения: 9.02.19).
4. Rajpurkar, P., Jia, R. and Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. [Электронный ресурс] // arXiv.org. URL: <https://arxiv.org/abs/1806.03822> (дата обращения: 9.02.19).
5. Sun, F., Li, L., Qiu, X. and Liu, Y. (2018). U-Net: Machine Reading Comprehension with Unanswerable Questions. [Электронный ресурс] // arXiv.org. URL: <https://arxiv.org/abs/1810.06638> (дата обращения: 9.02.19).
6. Vinyals, O., Fortunato, M. and Jaitly, N. (2015). Pointer Networks. [Электронный ресурс] // arXiv.org. URL: <https://arxiv.org/abs/1506.03134> 1711.07341 (дата обращения: 9.02.19).



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

Москва, 2019