MIPT DUDES

Презентация решения задачи хакатона

Содержание презентации

Постановка задачи и анализ ее особенностей

Постановка

«На основе открытого датасета научить голосового помощника наиболее точно отвечать на вопросы пользователей...»

Важные аспекты:

- Задача ранжирования
- Оцениваемой метрики нет

Проблемы которые можем предвидеть:

- Малый объем данных для обучения, что значительно увеличивает вероятность переобучения
- «Грязный» датасет – важна обработка
- Выбор целевого показателя – открытый вопрос для команды

Baseline решение

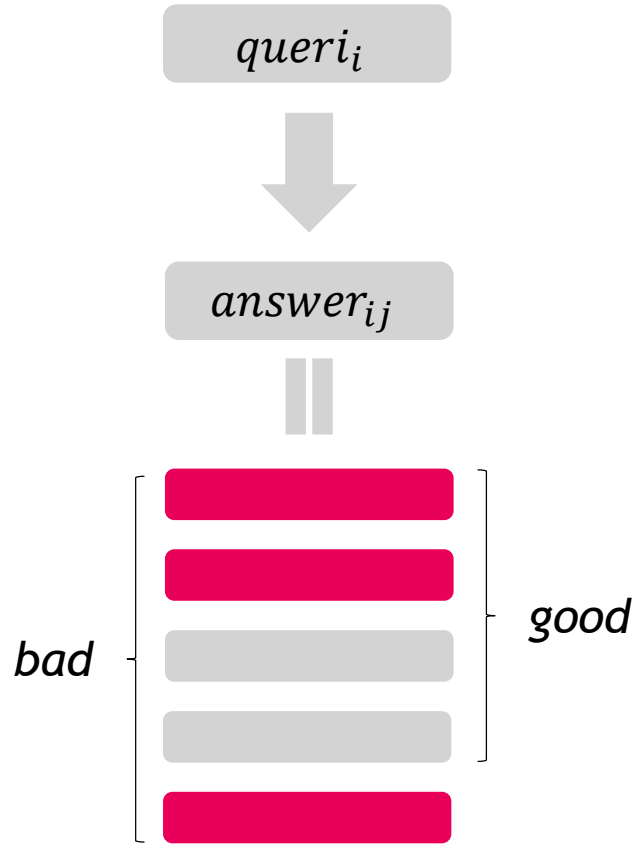
Постановка задачи

- Имеем датасет с вопросами и возможными ответами на них. Каждый ответ имеет метку - правильный он или нет.

$$D = \{queri_i, answer_{ij}, label_{ij}\}_{i,j}^{n,m}$$

- Решаем задачу ранжирования ответов на вопрос:

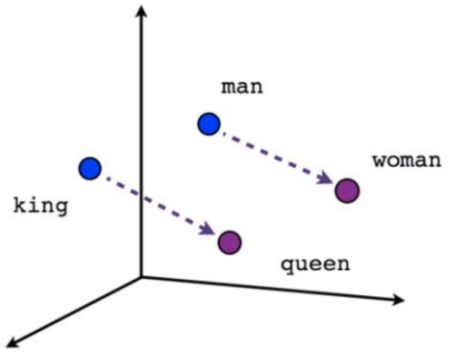
  для каждого блока { вопрос + возможные ответы } - должны быть отсортированы ответы в порядке – сначала все правильные, потом все неправильные.



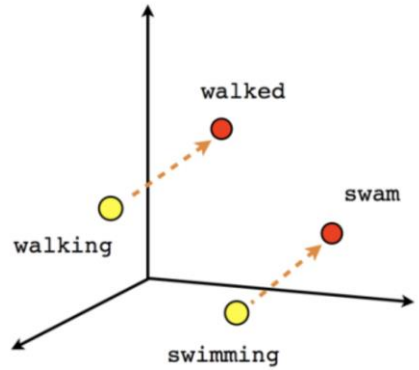
- Модель должна возвращать отсортированные датасет с указанием уверенности в ответах. Дальнейшей обработкой занимаются на следующем этапе.

Модель - bi-encoder

- Переводим слова в векторное представление – эмбеддинги
- Модель считает косинусную меру близости между вопросом и ответом
- Ранжируем ответы по полученным значениям*



Male-Female

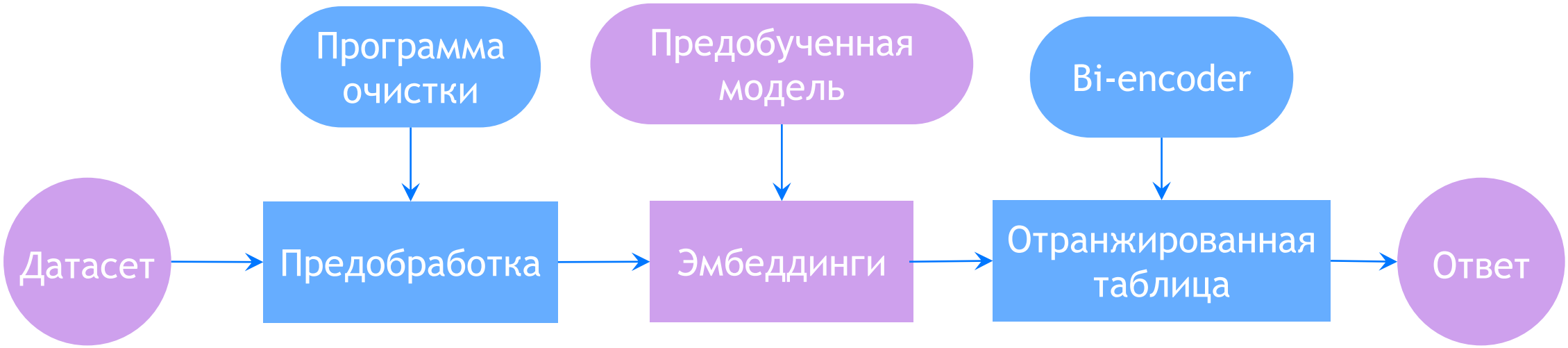


Verb tense

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

**брался softmax по значениям для приведения к значениям уверенности модели в ответе*

Pipe-line базового решения



Предобработка

- анализ данных
- очистка текста
- лемматизация текста
- …

Выбор модели и получение эмбеддингов

Модель

- Подсчёт расстояний
- Сортировка
- Ранжирование
- Выбор лучших

Предобработка

Предобработка состоит из нескольких этапов:

- Небольшой EDA
- Удаление стоп-слов с помощью библиотеки stop-words
- Лемматизация текста при помощи pymorphy3
- Приведение текста к нижнему регистру
- Удаление лишних знаков препинания
- Замена гласных с ударением
- Удаление текстов состоящих более чем из 6000 символов
- Токенизация

Эмбеддинги

- Задача относится к Intent classification (IC)
- На основе лидерборда эмбеддингов были протестированы : MUSE-3, DeepPavlov, deepvk, rubert-tiny2, LaBSE-en-ru.
- Лучше всего себя показал : deepvk

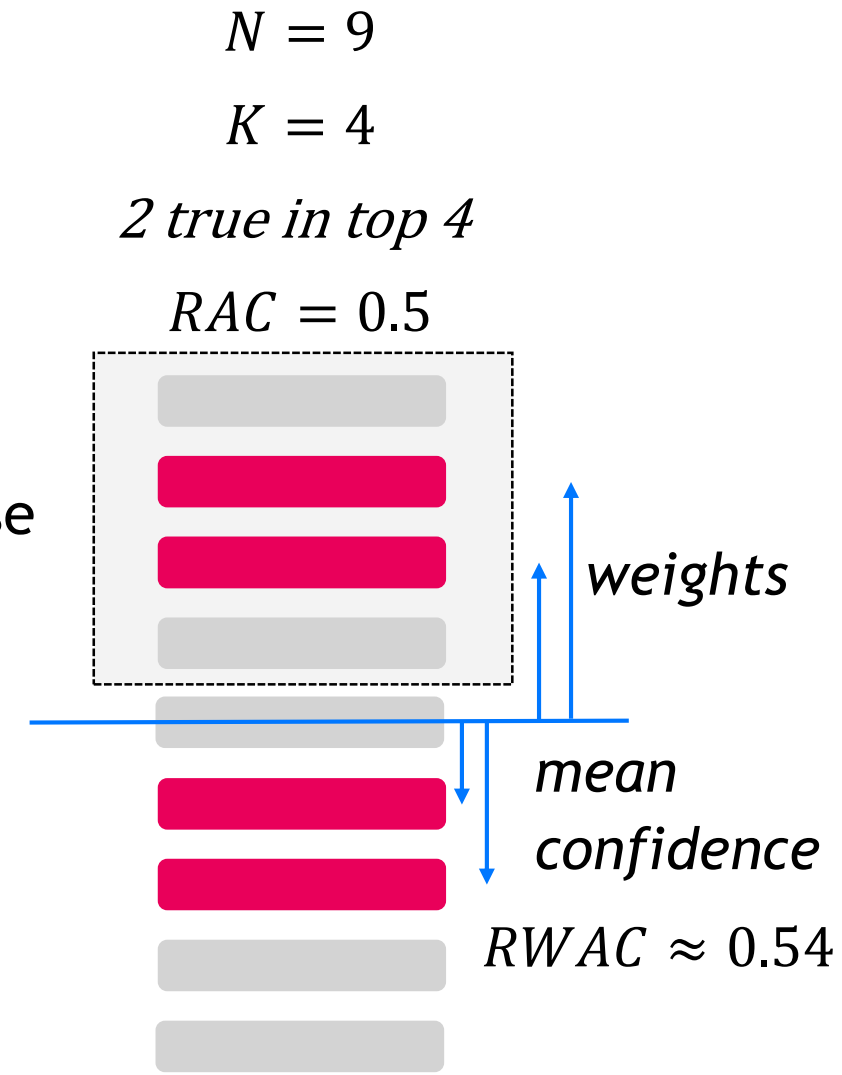
Метрики и результаты

- **NDSG** - *переменное число ответов*
- **MRR** (Mean Reciprocal Rank) - *слишком позитивная*
- **Presign@k** -*у нас переменное число ответов*

Исходя из идей вышеперечисленных, придуманы эти две метрики:

- Метрика покрытия релевантных ответов (RAC)
- Метрика покрытия с весами уверенности (RWAC)

$$\mathrm{RAC} = \frac{\left(\sum_i^{max(3,K)} label_i\right)}{K}$$



Результаты	**MRR**	**RAC**	**RWAC**
Baseline model	1.5	0.82	0.88

Продвинутое решение v1

Постановка задачи

- Сведём задачу к задаче классификации:
  Пара вопрос ответ -> уверенность модели в том что ответ подходит под вопрос

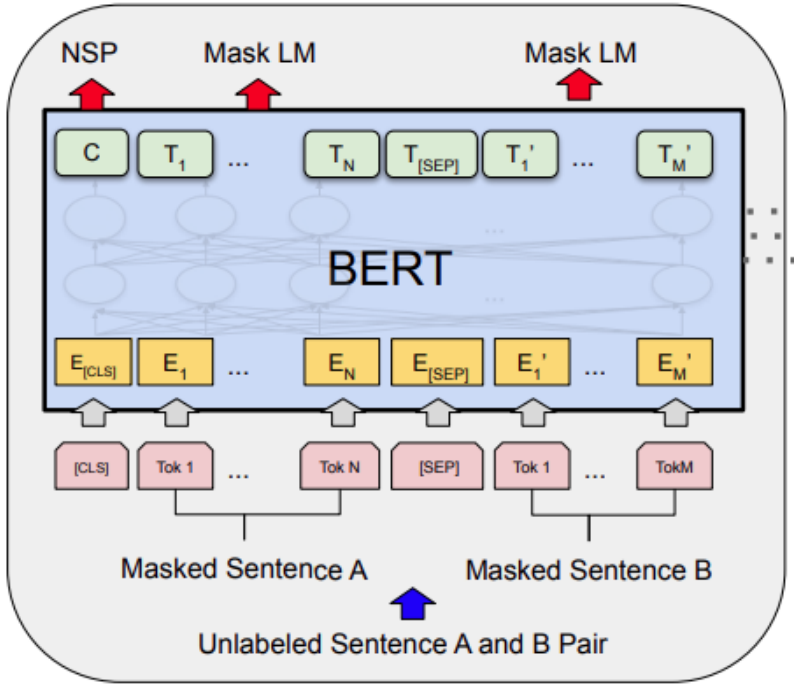
- Для решение исходной задачи:
  - Группируем по запросам
  - В итоговой таблице получаем результат работы модели для каждой пары запрос-ответ
  - Берём softmax от полученных значений и ранжируем по нему

- Метрики ранжирования можем оставлять как и в baseline

Query	Answer	Prob
text	*ans 1*	*0.30*
text	*ans 2*	*0.28*
text	*ans 3*	*0.22*
text	*ans 4*	*0.20*

Модель и результаты

- Используемая модель - `rubert-tiny`.
  Идея в использовании BERT-подобной архитектуры



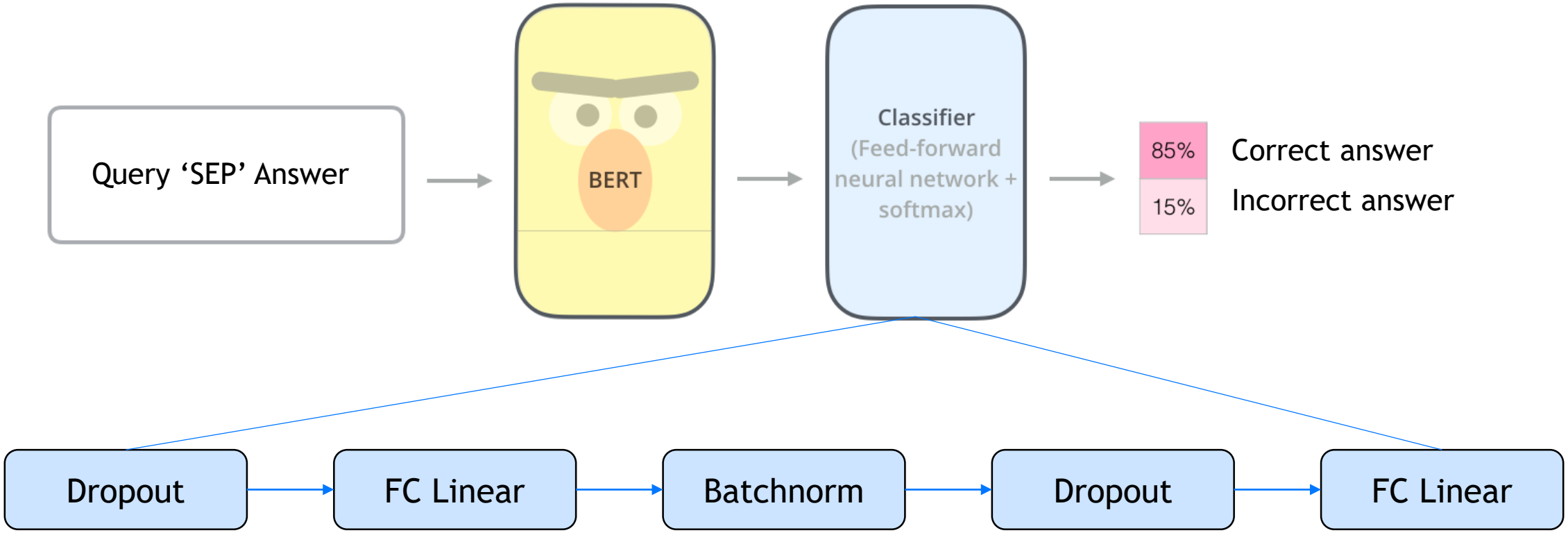
Результаты	MRR	RAC	RWAC
Baseline model	1.5	0.82	0.88
v1 model	3.27	0.510	0.514

- **Анализ результатов:**
  - Взята маленькая модель и она не чувствует специфичность вопросов
  - Не используются наши хорошие эмбеддинги, так что не удаётся добавить знания в эту модель. Хочется дообучить на нашем датасете
- **Идея улучшения** – взять за основу BERT и дообучить модель, добавив несколько линейных слоёв перед выходом для обучения классификации.

Продвинутое решение v2

Эпизод IV: Новая надежда

- Постановка задачи классификации и превращение её в необходимое ранжирование остаётся таким же



Обучение и результаты

- Обучение состояло из 2 этапов: обучение с «замороженным» BERT-ом, дообучение всей модели
- Были опробованы разные архитектуры выходной NN. Общая проблема для всех моделей – серьезное переобучение. Стандартные методы борьбы с ним(регуляризация, drop-out-ы и тд) смягчили этот эффект, но не сильно . Мы полагаем, что такое поведение обусловлено сильной специфичностью данных
- Ограниченность вычислительных ресурсов не позволила использовать большие модели.
- Итоговые метрики

Результаты	**MRR**	**RAC**	**RWAC**
v2 model	2.5	0.62	0.90

Анализ полученных результатов

Полученные результаты

Baseline решение

Bi-encoder на косинусном расстоянии с хорошими эмбеддингами

Up-grade v1 решение

Сведение задачи к классификации, использование модели BERT

Up-grade v2 решение

Добавление «головы» к предыдущей модели, дообучение BERT-а на нашем датасете

Результаты	MRR	RAC	RWAC
Baseline model	***1.51***	***0.82***	0.88
v1 model	3.27	0.51	0.51
v2 model	2.52	0.62	***0.90***

Итоговая таблица метрик для каждой модели

Summary : baseline решение дает неплохую точность ответов, использовав малые вычислительные ресурсы. Модели использующие BERT не сильно, но все-таки выигрывают baseline. Мы полагаем, что наше решение имеет потенциал к развитию.

Планы улучшения

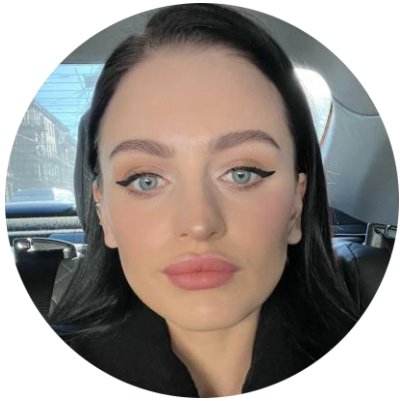
- Проблема с дообучением модели:
  - Увеличение размера выборки.
  - Использование другой bert-подобной модели в архитектуре.
- Предобработка ввода в модель
  - Поиск релевантных ответов среди всех ответов на все вопросы.
  - Поиск похожих вопросов, при нестандартном вводе
- Постобработка
  - Глубокий анализ топ-3 ответов.
  - Выделение чёткого ответа на поставленный вопрос.

Команда

Сергей Фирсов

Матвей Кервинен

Даниэль Сахаров



Татьяна Бушуева

Вадим Касюк

References

References

- Курс ЦК «Продвинутые методы машинного обучения»
- Курс лекций по машинному обучению Воронцова К.В.
- Курс лекций по машинному обучению и NLP Нейчева Р.Г и Гончаренко В.В.
- Авторы портала «Хабр» :
  - Практические аспекты ранжирования ответов виртуального ассистента Салют
  - Cross-Encoder для улучшения RAG на русском
- Лидербоард эмбедингов по разным задачам для русского языка :
  https://github.com/avidale/encodechka