Московский государственный технический университет
имени Н.Э. Баумана (национальный исследовательский университет)

Разработка чат-бота помощника для студентов вуза с использованием RAG

Студент группы ФН12-71Б: Н.Ю.Яшин
Научный руководитель: Д.Н.Федянин
Консультант: О.С.Ткачёва

Москва, 2024 г.

Постановка задачи

Цель работы:

Исследовать принципы построения Retrieval-Augmented Generation систем и создать прототип чат-бота.

Задачи:

- Собрать и подготовить данные для задачи.
- Создать модель Retrieval
- Реализовать Augmentated Generation.
- Оценить качество моделей.

Объекты в Natural Language Processing

Множество $\mathcal{V} = \{t_1, t_2, \dots t_N\}$ назовём **словарём**, а его элементы **токенами**.

Элементы множества $\mathcal{T} = \{(t_1, t_2, \dots, t_k) \mid t_i \in \mathcal{V}, \quad k \geq 1\}$ назовём **текстами**.

Определим множество **документов** - $D \subset \mathcal{T}$.
$\mathcal{D} = \{(t_1, t_2, \dots, t_k) \mid t_i \in \mathcal{V}, \quad k \geq n\}$

Пусть $d = (t_1, t_2, \dots, t_N)$ - документ.
Рассмотрим $ch_i = (t_{(i-1)m+1}, t_{(i-1)m+2}, \dots, t_{im})$.
Множество $C_d = \{ch_i \mid i = 1, 2, \dots, \frac{N}{m}\}$ назовём множеством **чанков** документа d.

Large Language Model

Большие языковые модели работают в два этапа:

1. Строится распределение вероятностей для токенов из словаря: $P(t|t_1, t_2, \dots t_k, \theta)$
2. Далее токен t_{k+1} выбирается случайно: $t_{k+1} \sim P(t|t_1, \dots t_k,), t \in V$

Большую языковую модель можно рассматривать как функцию
$LLM : T \to V$
$t_{k+1} = LLM(t_1, t_2, \dots, t_k)$

Пример использования:

$$\text{LLM}(\texttt{'сегодня'}, \texttt{'я'}, \texttt{'ходил'}, \texttt{'на'}) = \texttt{'работу'}$$

Large Language Model

Результат можно продолжать рекурсивно:
$t_{k+2} = LLM(t_1, \ldots, t_k, LLM(t_1, \ldots, t_k))$

Так мы получаем Instructive LLM:
$answer = ILLM(prompt)$, где $answer \in \mathcal{T}$, $prompt \in \mathcal{T}$.

Augmented-Generation: $AugLLM : \mathcal{T}^2 \to \mathcal{T}$
$AugLLM(prompt, context) = answer$

Сбор данных

В качестве документов я взял три сайта:

1. https://iu5bmstu.ru/index.php?title=Учреждения
2. https://ru.wikipedia.org/wiki/Московский_государственный_технический_университет_имени_Н._Э._Баумана
3. https://park.vk.company/faq/faq-header-18

Суммарно получилось собрать 74 чанков.

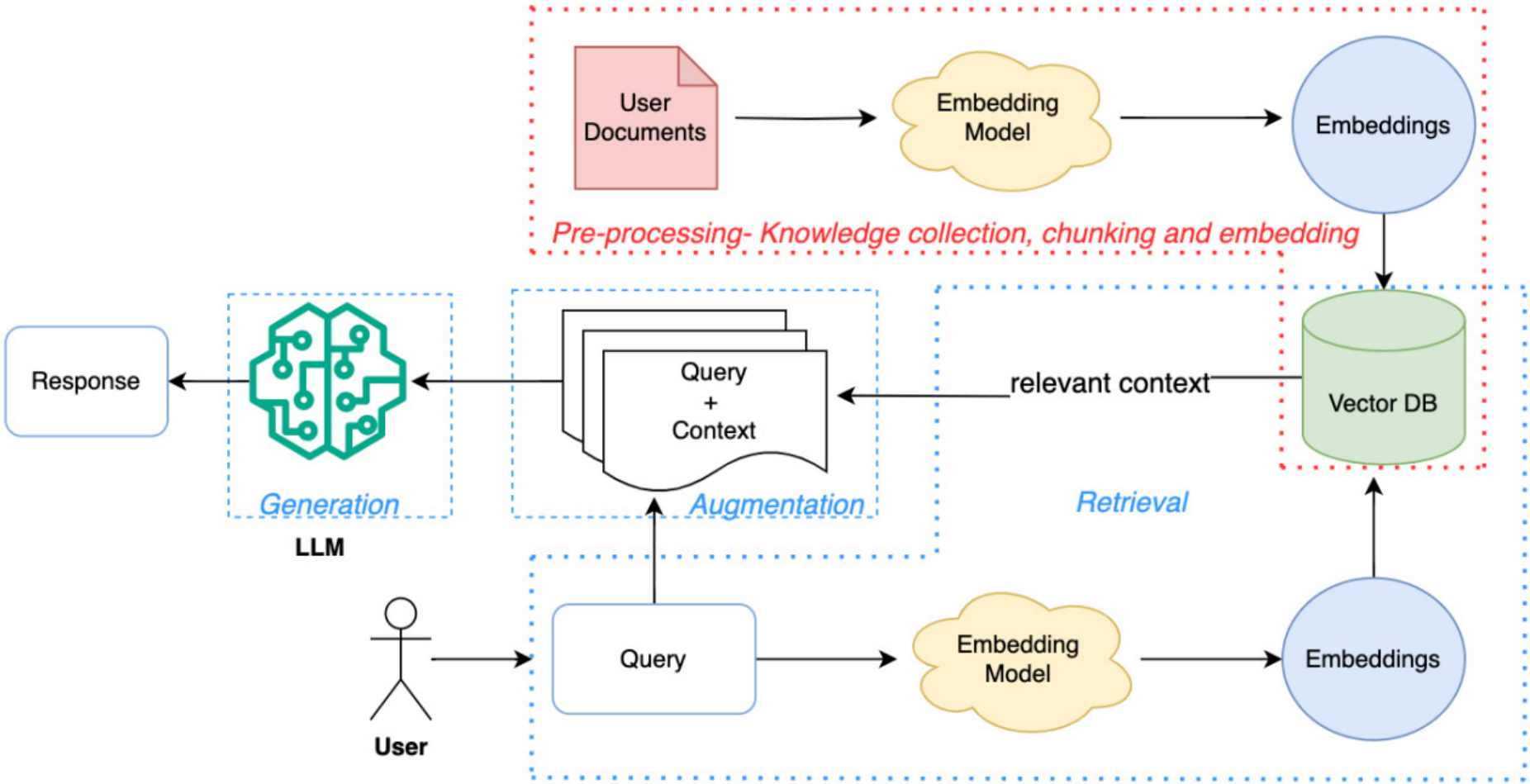
Для каждого чанка необходимо составить релевантный вопрос и ответ на этот вопрос. Для этого я использовал синтетические данные, полученные с помощью YandexGPT.

Пример данных

Пример

chunk: «Размер стипендии составляет 15 000 рублей в месяц- Стипендия от VK начисляется ежемесячно в течении 4 месяцев (сентябрь - декабрь 2023 для осеннего семестра; февраль 2024 - май 2024 для весеннего семестра)»,
question: «Сколько месяцев будет выплачиваться стипендия от VK и какой размер этой стипендии? »,
answer: «Стипендия от VK начисляется ежемесячно в течение 4 месяцев: с сентября по декабрь 2023 года для осеннего семестра и с февраля по май 2024 года для весеннего семестра. Размер стипендии составляет 15 000 рублей в месяц.»

Архитектура RAG (Retrieval Augmentation-Generation)



Архитектура модели Bidirectional Encoder Representations from Transformers

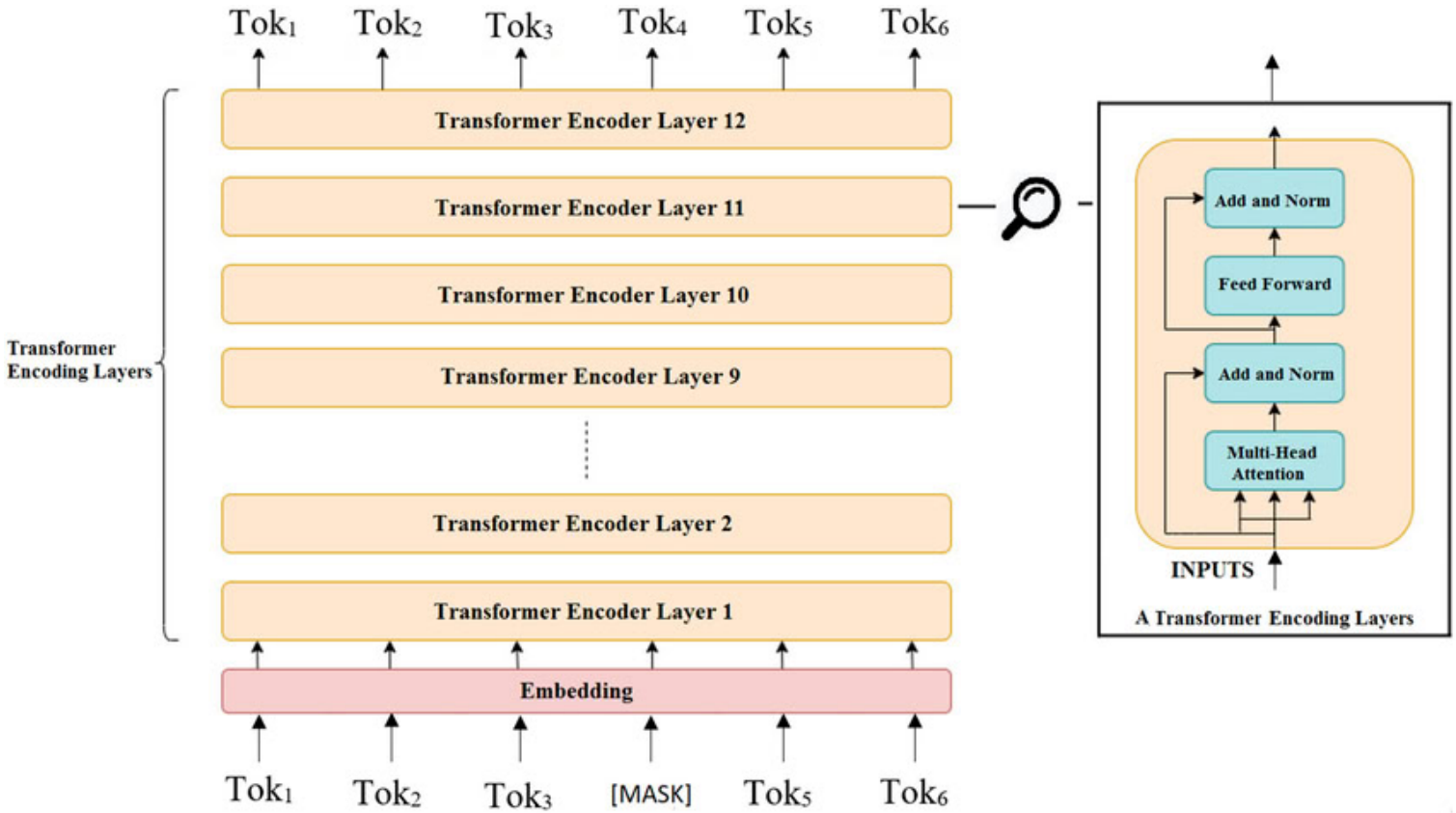


Рис.: Архитектура BERT

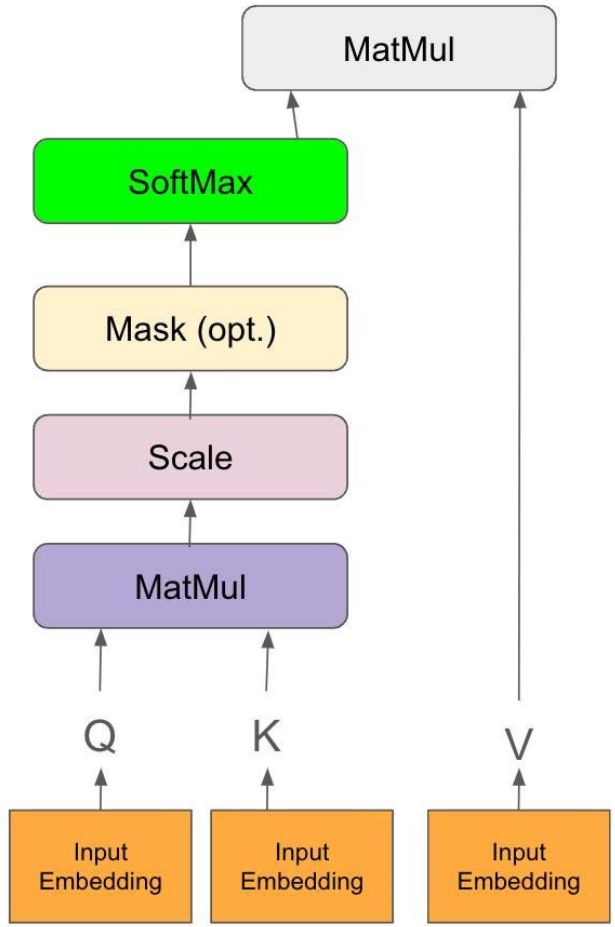
Механизм внимания

На вход подаётся m векторов X_i. Каждый вектор отвечает за свой токен.

Эти вектора преобразуются с помощью матриц W_Q, W_K, W_V.
$Q_i = W_Q X_i \; K_i = W_K X_i \; V_i = W_V X_i$.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

где $softmax(x_1, \ldots, x_n)_i = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}}$



Модель Retrieval

Получаем, что модель BERT - это функция, которая преобразует текст в вектор: $\text{BERT} : \mathcal{T} \to \mathbb{R}^m$

Полученные вектора называются **эмбеддингами**.

И таким образом, Retrieval модель работает следующим образом:

$$Retrieval(prompt, C) = \arg\min_{ch \in C} CosDist(\text{BERT}(prompt), \text{BERT}(ch))$$

где $CosDist(a, b) = 1 - \frac{(a,b)}{\|a\|\|b\|}$.

Оценка модели Retrieval

Метрики:

$$\text{Recall@k} = \frac{\text{Количество релевантных чанков среди топ-k}}{\text{Общее количество релевантных чанков в датасете}}$$

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q}$$

$$\text{MR} = \frac{1}{|Q|} \sum_{q \in Q} \text{rank}_q$$

здесь rank_q означает номер релевантного чанка q в выдаче

$$\text{NDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}}$$

где

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$$

$$\text{IDCG@k} = \sum_{i=1}^{k} \frac{1}{\log_2(i+1)}.$$

Полученные метрики

После тестирования на 74 вопросах из датасета, получились следующие значения метрик:

- **Средний Recall@1**: 0.638
- **Средний Recall@3**: 0.809
- **Средний Recall@5**: 0.851
- **Средний MRR**: 0.731
- **Средний Mean Rank**: 3.064
- **Средний NDCG@1**: 0.638
- **Средний NDCG@3**: 0.732
- **Средний NDCG@5**: 0.749

Augmented Generation

В качестве LLM я взял предобученную модель на 1 миллиард параметров: **Vikhr-Llama-3.2-1B-instruct**.
Модель я взял с ресурса *Hugging Face*.
Для того, чтобы такая модель умела понимать чётко разделять контекст и вопрос, данные надо подавать определённого вида.

Структура текстовых данных

<|begin_of_text|><|start_header_id|>system<|end_header_id|>
<|eot_id|><|start_header_id|>user<|end_header_id|>
context
<|eot_id|><|start_header_id|>user<|end_header_id|>
question<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
answer
<|eot_id|>

Демонстрация работы системы

Вопрос	RAG Система	ChatGPT
Какие языки программирования изучают на первом семестре в VK технопарке?	На первом семестре в VK технопарке, изучаются C++ и Python.	В VK Технопарке студенты изучают различные языки программирования в зависимости от выбранной образовательной программы. Например, в подготовительном курсе по C++ рассматриваются основы работы с памятью, обработка исключений и базовые навыки обобщённого программирования. В рамках курса «Языки веб-программирования» от МГТУ им. Н.Э. Баумана используется язык Ruby. Таким образом, выбор языка программирования зависит от конкретной программы обучения в VK Технопарке.
Сколько месяцев будет выплачиваться стипендия от VK и какой размер стипендии?	Стипендия от VK выплачивается ежемесячно в течение 4 месяцев. Стоимость стипендии составляет 15 000 рублей в месяц.	В VK Технопарке предусмотрены стипендии для студентов, однако точная информация о размере и продолжительности выплат не указана в доступных источниках. Рекомендуется обратиться напрямую в VK Технопарк для получения актуальной информации о стипендиях.

Оценка качества модели

Метрики:

$$\text{BLEU} = \min\left(1, \frac{C_n}{R_n}\right) \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\text{METEOR} = \frac{\text{Precision} \cdot \text{Recall}}{\alpha \cdot \text{Precision} + \beta \cdot \text{Recall} + \gamma \cdot \text{Penalty}}$$

$$\text{ROUGE-1} = \frac{\sum_{w \in \text{overlap}} \text{count}(w)}{\sum_{w \in \text{reference}} \text{count}(w)}$$

Усреднённые по всем вопросам результаты:

BLEU = 0.16, METEOR = 0.33, ROUGE-1 = 0.48

Завершение

Спасибо за внимание!