

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
ИНСТИТУТ ИНТЕЛЛЕКТУАЛЬНОЙ РОБОТОТЕХНИКИ

Разработка интерактивного учебного пособия с ответами на естественном языке на основе Retrieval Augmented Generation

Сыренний Илья Игоревич
Бакалавриат
Курс 4, группа 21930

Научный руководитель:
Оглезнев Никита Сергеевич,
сотрудник КафИСТИИР, ассистент

Новосибирск 2025

Введение

Цель работы:

Разработать систему, использующую подход Retrieval Augmented Generation (RAG) для помощи пользователям в изучении научных статей в формате PDF. Система должна отвечать на вопросы пользователя по загруженному документу, а также предоставлять релевантные фрагменты

Задачи:

Поиск и анализ литературы в рамках изучения предметной области

Проектирование и разработка системы

Разработка способов оценки системы

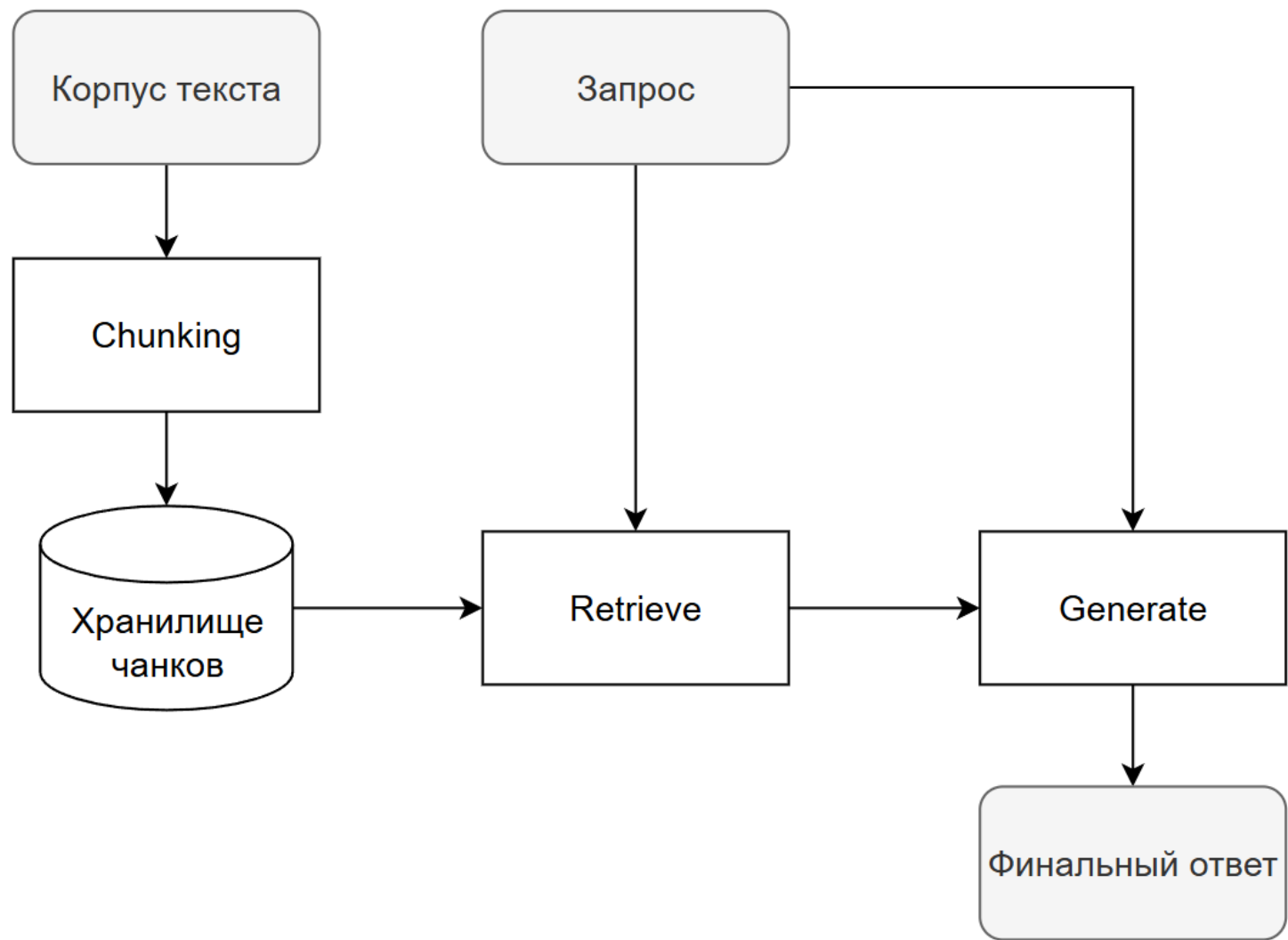
Retrieval Augmented Generation

Проблема:

Большие текстовые массивы сложно использовать напрямую из-за ограниченного контекста моделей. Это затрудняет извлечение точной и актуальной информации

Решение:

RAG сочетает поиск данных и генерацию текста. Поисковая система (Retrieval) находит релевантные фрагменты текста из базы данных. Генерация (Generation) создает ответ, основываясь на найденных данных



Устройство RAG-пайплайна

Разработка системы

Проектирование:

Для проектирования системы были использованы UML-диаграммы, которые позволили структурировать требования, визуализировать архитектуру и глубже понять функциональность системы

Архитектура

Разработано веб-приложение в виде чата для общения с пользователем, пользователь может загружать документы, задавать вопросы, получать ответы в потоковом режиме

Поисковая система

Для поддержки мультязычного поиска реализован векторный поиск, учитывающий семантические особенности текста

Файлы

Attention.pdf

1 из 15

— | +

🔍 T 📄

Attention Is All You Need

Ashish Vaswani*
Google Brain
vaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszko
Google Reser
usz@google.

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.cc

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer model. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly all of the design decisions. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase. Llion also experimented with novel model variants, was responsible for our initial codebase, inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of the model, replacing our earlier codebase, greatly improving results and massively accelerating research.

[†]Work performed while at Google Brain.
[‡]Work performed while at Google Research.

Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

выполнения задачи.

В контексте нейронных сетей attention используется для выделения наиболее важных элементов входных данных, таких как слова или символы, и для вычисления представления последовательности. Это достигается за счет использования механизма само-внимания (self-attention), который позволяет модели relacionать разные позиции одной последовательности для вычисления ее представления.

Функция attention принимает на вход запрос, набор ключей и значений, и возвращает выходной вектор, который является взвешенной суммой значений, где веса определяются сходством между запросом и ключами. Это позволяет модели сосредоточиться на наиболее важных элементах входных данных и игнорировать менее важные.

В статье "Attention Is All You Need" авторы предлагают использовать механизм attention для нейронного машинного перевода, что позволяет модели более эффективно обрабатывать входные данные и улучшать качество перевода. Они также предлагают использовать много-головое внимание (Multi-Head Attention), которое позволяет модели одновременно учитывать несколько аспектов входных данных и улучшать качество представления последовательности.

Фрагменты

Спросите что-нибудь...

Пользовательский интерфейс веб-приложения

Работа с PDF-файлами

Индексация

Система загружает PDF-файлы и разбивает их на логические фрагменты в реальном времени. Для восстановления макета документа используется CRF-модель, обученная на разметке научных статей. Затем применяется алгоритм семантического чанкинга для формирования базы знаний

Скорость загрузки документов

Книга “Deep Learning” (Goodfellow et al., ~800 страниц) индексируется за ~7 минут

Научная статья (15 стр.) индексируется за ~2 секунды

Система оптимизирована для работы в интерактивном режиме: индексация начинается сразу после загрузки документа

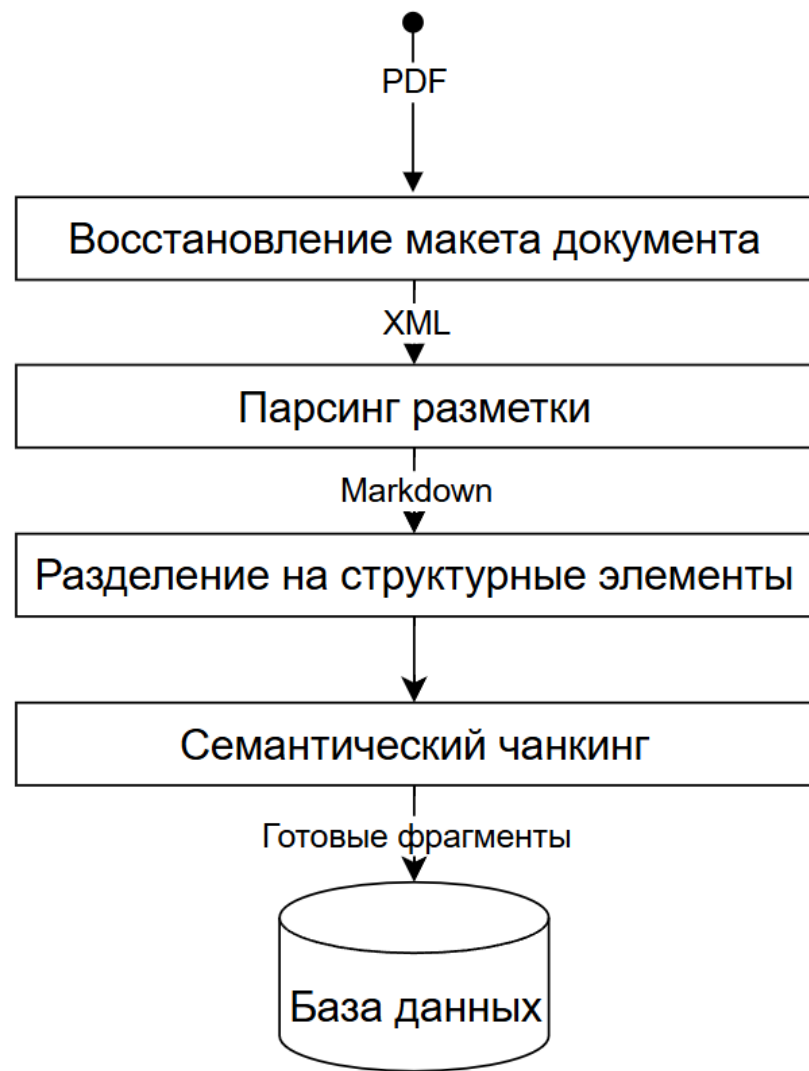


Схема работы этапа индексации

Оценка системы

Датасет: FRAMES

Содержит 824 вопроса, требующих информации из 2-15 статей из Wikipedia. Для каждого вопроса представлен эталонный ответ, а также список релевантных статей из Wikipedia

Метрики: RAGAS

Использованы метрики LLM-as-Judge для оценки качества генерации, а также релевантности найденных для ответа фрагментов

Пилотное тестирование

Приложение было развернуто и передано 6 студентам НГУ для сбора обратной связи о работе системы. Собрана обратная связь по качеству ответов, удобству интерфейса и скорости работы системы

Метрики системы

Сравнение метрик на разных конфигурациях

В таблице приведены метрики, рассчитанные на основе датасета FRAMES и метрик LLM-as-Judge из библиотеки RAGAS. Метрики рассчитаны на выборке из 50 вопросов и 330 статей из Wikipedia.

Наименование метрики	Agentic-RAG	RAG	LLM
Релевантность поиска	0.51	0.36	-
Достоверность генерации	0.68	0.54	-
Правильность генерации	0.62	0.48	0.26
Среднее время работы (с)	29	6.5	3

Метрики пилотного тестирования

Доверительные интервалы оценок пользователей

Поскольку выборка из 6 человек маленькая, были рассчитаны доверительные интервалы, которые помогли определить, в каком диапазоне с вероятностью 95 процентов лежат истинные значения для всей выборки.

Аспект приложения	Среднее	Нижняя оценка	Верхняя оценка
Точность ответов	4.16	2.93	5.39
Скорость работы	5.83	5.04	6.62
Удобство	4.5	3.39	5.6
Полезность	4.0	2.51	5.48
Общая оценка	4.83	3.6	6.06

Результаты и выводы

Использование RAG позволяет существенно повысить качество ответов, подключая внешние проверенные источники знаний.

Разработана модульная и масштабируемая RAG-система с возможностью индексации документов и генерации ответов в реальном времени.

Система успешно прошла пилотное испытание и готова к практическому применению.