004.021

Максименко Олег Евгеньевич

Аспирант НИЯУ МИФИ, Кафедра 22

115409, Россия, Москва, Каширское шоссе, 31

email: oleg.maksimenko@me.com

ОБЗОР РУССКОЯЗЫЧНЫХ ДАТАСЕТОВ ДЛЯ ИНФОРМАЦИОННОГО ПОИСКА

Аннотация: Статья посвящена обзору и сравнению датасетов для информационного поиска на русском языке. Задача информационного поиска начала решаться с данных на английском языке, однако оказалась востребована и для русского языка. Данные можно получать посредством перевода имеющихся текстов на английском языке, однако такой подход не всегда может быть релевантен из-за неучета языковых особенностей, в связи с чем и были созданы соответствующие текстовые наборы данных. Датасеты могут быть оценены по объему, качеству соответствия запросам и ответам на них, тематике с использованием различных метрик.

Ключевые слова: датасет, информационный поиск, метрики, ранжирование, объем

Abstract: The article is devoted to the review and comparison of datasets for information search in Russian. The task of information search began to be solved with data in English, but it turned out to be in demand for the Russian language. Data can be obtained by translating existing texts into English, but this approach may not always be relevant due to the lack of consideration of language features, and therefore the corresponding text datasets were created. Datasets can be evaluated by volume, quality of matching requests and responses to them, and subject matter using various metrics.

Keywords: dataset, information search, metrics, ranking, volume.

Введение

Информационный поиск в настоящее время актуален в связи с широким распространением информационных технологий и средств связи по всему миру. И хотя огромный объем информации изложен на английском языке, существует огромное количество материалов на русском. В информационном поиске используются как классические алгоритмы ранжирования, так и подходы основанные на машинном обучении [1].

Датасеты для информационного поиска на русском языке нужны по нескольким причинам. Во-первых они активно используются при разработке и тестировании алгоритмов ранжирования. Они позволяют разработчикам создавать, обучать и тестировать алгоритмы информационного поиска, адаптированные под особенности русского языка и соответствующие культурные контексты, что отражается на финальном качестве поиска и релевантности результатов для конечных пользователей.

Во-вторых, датасеты применяются при использовании алгоритмов, в основе которых лежит машинное обучение и часто представляют из себя большие объёмы размеченных данных. Конкретным примером может служить обучение больших языковых моделей, применение которых началось с 2010-х годов.

Также датасеты могут использоваться для оценки точности систем информационного поиска. Если предварительно разделить выборки на обучающие и тестовые, то можно сравнить результаты работы системы на с заранее известными ответами, и на основе метрик определить, насколько хорошо система справляется с задачами поиска.

Дополнительно можно отметить, что рассматриваемые датасеты могут помочь в изучении некоторых особенностей русского языка, таких как

морфология, синтаксис и семантика, что может быть использовано для разработки более эффективных алгоритмов поиска.

Материалы и методы

В статье будет рассмотрено 5 различных датасетов, отличающихся по содержанию и формату хранимых данных.

SberQuAD - это русскоязычный датасет , созданный по аналогии с популярным английским набором данных SQuAD (Stanford Question Answering Dataset). Создан компанией Сбербанк и содержит 50 тысяч записей, из которых 15 тысяч используются для обучения и 25 - для тестирования [2]. Он формировался на основе данных из Википедии содержит контекстные отрывки, вопросы к ним и соответствующие ответы. Между форматами SQuAD и SberQuAD есть два различия. Во-первых, SberQuAD не сообщает, к каким страницам Википедии относится абзац и во-вторых, каждый ответ представлен строкой без соответствующей начальной позиции в абзаце. Большинство вопросов в наборе данных начинаются либо с вопросительного слова, либо с предлога: что, в, как, кто, когда, где, сколько.

RIA-News - это датасет новостных статей, собранный на основе материалов информационного агентства РИА Новости.

Датасет содержит два основных компонента - это корпус текстов (704 тысячи записей) и набор запросов (10 тысяч записей). В нем содержатся новостные материалы за период с января 2010 по декабрь 2014 года. Всего в предоставленном наборе содержится 1003869 новостных статей со средним объемом заголовка 9,5 слов и текста 315,6 слов [3].

ruSciBench-retrieval - это специализированный датасет для оценки моделей информационного поиска в научной сфере на русском языке. Специальный набор научных текстов был создан с использованием

крупнейшей в России электронной библиотеки научных публикаций eLibrary. Из всех статей eLibrary были выбраны только объемные аннотации (более 50 символов). Все html-теги из текста были удалены, определен язык тезисов с помощью библиотеки lingua и удалены все статьи, в которых отсутствовали аннотации на русском или английском языках. Изначально было 197 220 статей, после фильтрации осталось только 182 264 [4]. Тематически данные охватывают гуманитарные, естественные, технические, медицинские науки и исследовании в области права. Датасет предназначен для:

- Оценки качества алгоритмов поиска в сфере науки и их тестирования
- Разработки систем поиска информации в научных текстах

ru-facts — это текстовый датасет на русском языке, содержащий разнообразные факты и новостные материалы. Средняя длина текста составляет 198 символов, минимальная - 10 символов, а максимальная - 3402 символа. Сам набор данных был сформирован с использованием трех основных подходов - тексты, сгенерированные с помощью модели перефразирования, перевода датасета для проверки фактов и дополнения существующего текста.

В датасете представлены материалы по следующей тематике:

- Новостные события
- Исторические факты
- Научные данные

rus-trec-covid - это русскоязычный датасет, созданный для оценки систем информационного поиска в области медицинских и научных исследований, связанных с COVID-19. Может быть использован для создания эффективных систем поиска в научной и медицинской сфере.

Сам датасет сформирован из медицинских исследований, статей, посвященных респираторным заболеваниям, данных клинических наблюдений и описаний медицинских систем.

Результаты

Рассматриваемые датасеты можно оценивать по многим параметрам, некоторые из которых - объем, формат хранения. При выборе датасета важно учитывать комбинацию этих критериев в зависимости от конкретной задачи и целей исследования. Некоторые критерии могут быть более важными для определенных типов задач, поэтому необходимо проводить приоритизацию в соответствии с потребностями проекта. Стоит опираться на количественные, качественные характеристики, тематические аспекты, практические критерии. Также следует обратить внимание на год создания датасета, так как это может быть важно в контексте работы с актуальными данными.

Датасет	Объем	Формат	Создание
SberQuAD	50000 записей	CSV	2019 год
RIA-News	700000 записей	JSON	2010 – 2014 год
ruSciBench-retrieval	200000 записей	JSON	2021 год
ru-facts	10000 записей	JSON	2023 год
rus-trec-covid	171000 записей	JSON	2021 год

Таблица 1 - Характеристики датасетов

Также для оценки состава и релевантности данных была применена метрика BM25, часто используемая для ранжирования результатов, выдаваемых по запросу. BM25 —семейство функций ранжирования документов, которые оценивают число ключевых запросов в каждом из документов [5].

Данные из датасетов перед оценкой прошли несколько этапов предварительной обработки:

- Лемматизация с помощью библиотеки Natural Language Tooolkit
- Очистка от мусорных слов с помощью заранее заготовленного словаря из библиотеки Natural Language Tooolkit
- Очистка от знаков препинания и цифр

Датасет	BM25
SberQuAD	0.68
RIA-News	0.63
ruSciBench-retrieval	0.36
ru-facts	0.92
rus-trec-covid	0.44

Таблица 2 - Результаты оценки по BM25

Обсуждение

Результаты показывают, что при выборе датасета можно сделать следующие рекомендации для использования.

В случае реализации систем информационного поиска, не завязанных на узкую тематику, можно прибегнуть к RIA-News или ru-facts. ru-facts также хорошо подходит для тестирования базовых алгоритмов ранжирования за счет того, что его объем ограничен. Однако датасет из Риа-новостей может содержать устаревший материал.

Для реализации QA систем стоит обратиться к SberQuAD.

rus-trec-covid может быть использован для медицинского поиска или тестирования систем фильтрации медицинской информации, однако его применение строго ограничено заданной тематикой и может быть полезно

лишь части исследователей. Из всех представленных датасетов обладает самой узкой областью применения.

ruSciBench-retrieval может быть использован для реализации информационного поиска в сфере научных исследований. Также для оценки качества представленных в нем данных требуется наличие определенной квалификации.

Заключение

Был произведен обзор датасетов на русском языке, используемых для информационного поиска. Взяты 5 наборов данных разной тематики и годов. Объемы каждого подобного датасета составляют не менее нескольких десятков тысяч записей, что позволяет эффективно использовать их в рамках проведения экспериментов по части информационного поиска. Тематики текстов охватывают различные направления, начиная научными исследованиями и заканчивая новостными данными за определенные года. Разбиения по тематикам дает возможность конкретизировать область информационного поиска и благодаря преобразованиям над текстом реализовать различные системы информационного поиска.

Список литературы

1. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011
2. Павел Емов, Андрей Черток, Леонид Бойков, Павел Бреславский - Набор данных о понимании прочитанного на русском языке: описание и анализ, 2020 год
3. Даниил Гаврилов, Павел Каладин, Валентин Малых, Внимательная к себе модель для создания заголовков, 2019

4. А. Ватолин, Н. Герасименко, А. Янина и К. Воронцов, Rustic Bench: Открытый тест для представления научных документов на русском и английском языках, 2025
5. Л.В., Ю., К. Чжай, когда документы очень длинные, BM25 выходит из строя! SIGIR, 2011, с. 1103-1104

Referents

1. Manning K., Raghavan P., Schutze H. Introduction to information search. — Williams, 2011
2. Pavel Efimov, Andrey Chertok, Leonid Boytsov, Pavel Braslavsky - Russian Reading Comprehension Dataset: Description and Analysis, 2020
3. Daniil Gavrilov, Pavel Kaladin, Valentin Malykh, Self-Attentive Model for Headline Generation, 2019
4. A. Vatolin, N. Gerasimenko, A. Ianina & K. Vorontsov, Rustic Bench: Open Benchmark for Russian and English Scientific Document Representations, 2025
5. LV, Yu, K. Zhai, when the documents are very long, the BM25 fails! SIGIR, 2011, pp. 1103-1104