

Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М.В.ЛОМОНОСОВА»
(МГУ)

Научно-исследовательский вычислительный центр

УДК
004.9
Рег. № НИОКТР
125032004284-7
Рег. № ИКРБС

УТВЕРЖДАЮ
Директор
Научно-исследовательский
вычислительный центр,
д.ф.-м.н., проф., член-корр. РАН

_____ В.В. Воеводин

«_____» _____ Г.

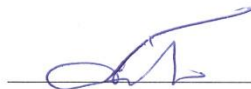
ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по теме:

Семантический поиск на основе нейросетевых моделей
информационного-поиска для русского языка: создание бенчмарка и
разработка новых моделей
(промежуточный)

Руководитель НИР:

заведующий лабораторией,
кандидат
физико-математических наук,



Добров Б.В.

Москва 2025

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель НИР,
заведующий ла-
бораторией, кан-
дидат физико-
математических наук



Добров Б.В.
(все разделы)

Исполнители:

стажёр-
исследователь



Ковалев Г.П.
(все разделы)

ведущий науч-
ный сотрудник,
кандидат физико-
математических наук,
доктор технических
наук



Лукашевич Н.В.
(все разделы)

старший науч-
ный сотрудник,
кандидат физико-
математических наук



Тихомиров М.М.
(все разделы)

ВВЕДЕНИЕ

Научной проблемой, рассматриваемой в проекте, является семантический информационный поиск, т.е. нахождение наиболее релевантных к запросу документов не только по совпадающим словам, но и по смыслу. Семантический поиск должен обеспечивать нахождение релевантных документов даже в условиях отсутствия лексического совпадения между запросом и документом. Появившиеся за последние пять лет методы и модели на основе нейросетевой архитектуры трансформер дают новые возможности для обучения специализированных моделей поиска, которые позволяют искать релевантные документы независимо от лексического выражения смысла; большие языковые модели позволяют применять новые подходы к расширению запросов и документов. Однако большинство подходов тестируются на данных для английского языка, оцениваются на англоязычных бенчмарках, качество работы на разных типах русскоязычных текстовых данных неизвестно.

Данный проект посвящен созданию бенчмарка, т.е. совокупности датасетов для тестирования моделей информационного поиска для русского языка, а также тестированию моделей на созданном бенчмарке. На текущий момент созданный бенчмарк содержит 27 датасетов различного происхождения, включая переведенные датасеты, датасеты из существующих мультязычных наборов данных, существующие русскоязычные датасеты, а также набор на основе созданных на основе интересных фактов Википедии.

ОСНОВНАЯ ЧАСТЬ

1. Для тестирования моделей информационного поиска на русском языке был создан бенчмарк RusBEIR. Модели тестируются в формате zero-shot без дообучения на датасетах.

Бенчмарк состоит из датасетов разного происхождения. Первая группа датасетов была автоматически переведена с датасетов на английском языке. Это важно для сопоставимости с результатами, полученными на других бенчмарках. Для этого было выполнено тестирование трех систем машинного перевода: Google.translate, Яндекс.переводчик и нейросетевая модель перевода nllb. Переводы этих моделей были протестированы двумя лингвистами. Лучшие оценки были получены Яндекс.переводчиком, Google.translate получил немного меньшие оценки. В результате для перевода датасетов был выбран

Google.translate, поскольку давал более удобные условия для доступа.

Вторая группа датасетов была взята из существующих мультязычных датасетов (mMARCO, Miracl, XQuAD, Tydi QA). Третья группа датасетов была взята из существующих русскоязычных датасетов (Ria-News, SberQuAD, ruSciBench, ru-facts).

Четвертая группа датасетов была создана на основе данных Википедии (см. следующий п.2).

В настоящее время бенчмарк RusBEIR содержит 27 наборов данных.

2. Датасеты из Википедии созданы на основе блока интересных фактов Википедии «Знаете ли Вы что?». Эта секция содержит факты, которые описаны в

Википедии. Они создаются авторами статей, при этом исходные предложения могут быть существенно перефразированы. Пример такого факта: «Ожерелье, заказанное Наполеоном для второй жены, следовало моде на «реки из бриллиантов». Такие интересные факты могут быть рассмотрены как запросы к данным Википедии для проверки подтверждения фактов (Fact checking). Каждый такой факт содержит ссылки на статьи из Википедии, жирным шрифтом выделена статья, в которой содержится ответ. Так, в приведенном примере есть выделенная ссылка на словосочетание «реки из бриллиантов», которая ведет на статью «Ривьера (ожерелье)». Таким образом, уже исходно имеется разметка для организации поиска по статьям Википедии.

Для 5500 фактов Википедии была выполнена разметка по предложениям. Предложения размечались по трехбалльной шкале. Предложение получало оценку 2,

если оно содержало факт полностью, и оценку 1, если оно содержало часть факта. В результате, на основе размеченных предложений были созданы датасеты из фрагментов статей разной величины: абзацы, а также фрагменты размером от 2 до 6 предложений. Таким образом, на основе интересных фактов Википедии были созданы следующие датасеты (для одних и тех же запросов):

- wikifacts-articles – полные статьи,
- wikifacts-sents – отдельные предложения,
- wikifacts-para – абзацы.
- wikifacts-window2, wikifacts-window3, wikifacts-window4, wikifacts-window5, wikifacts-window6 – фрагменты исходных статей разной величины по коли-

честву предложений, создаются путем сдвига по предложениям статей заданной совокупности предложений.

Абзацы и фрагменты получают оценку релевантности, равную максимальной оценке содержащихся в них предложений. Такой набор датасетов разной величины может использоваться для тестирования влияния размера документа на результаты различных моделей.

3. На созданном бенчмарке были протестированы: лексическая модель BM25, а также различные нейросетевые модели. Модель BM25 применялась к лемматизированным русским текстам (использовался анализатор PyMorphu3), было еще раз подтверждено, что результаты лексических моделей, применяемых к лемматизированным текстам, существенно выше, чем в применении к исходным текстам, из-за многообразия морфологических форм в русском языке.

В состав исследованных нейросетевых моделей входили кросс-энкодеры и биэнкодеры. Дополнительно использовался реранкер, который применялся для переранжирования 100 первых результатов первого этапа. В качестве реранкера была выбрана модель bge-reranker-v2-m3.

4. Создан гитхаб <https://github.com/kaengreg/rusBeIR> с набором удобных инструментов для применения и тестирования моделей на бенчмарке. Датасеты с разметкой релевантности опубликованы на сайте huggingface: <https://huggingface.co/collections/kaengreg/rusbeir>.

5. Создан небольшой датасет для тестирования методов RAG на русском языке для тестирования порождения ответов генеративными языковыми моделями. Датасет создан на основе вопросно-ответных датасетов из Rus BEIR. Для лучшего

сопоставления ответов моделей имеющиеся ответы дополнены вариантами и синонимами. Датасет опубликован на платформе huggingface (<https://huggingface.co/collections/bearberry/rusrag-benchmark>)

6. Проведена оценка ответов генеративных моделей на созданном RAG бенчмарке.

7. Опубликованы статьи и представлены доклады на международных и российских конференциях.

В текущем году была опубликована в материалах конференции Диалог-2025 статья

Building russian benchmark for evaluation of information retrieval models / G. Kovalev, M. Tikhomirov, E. Kozhevnikov et al. // In Proceedings of the International Conference "Dialogue-2025. — 2025. (индексируется в Скопусе)

По результатам участия в конференции DAMDID-2025 две статьи рекомендованы к публикации в журналах.

ЗАКЛЮЧЕНИЕ

Научной проблемой, рассматриваемой в проекте, является семантический информационный поиск, т.е. нахождение наиболее релевантных к запросу документов не только по совпадающим словам, но и по смыслу. Семантический поиск должен обеспечивать нахождение релевантных документов даже в условиях отсутствия лексического совпадения между запросом и документом. Появившиеся за последние пять лет методы и модели на основе нейросетевой архитектуры трансформер дают новые возможности для обучения специализированных моделей поиска, которые позволяют искать релевантные документы независимо от лексического выражения смысла; большие языковые модели позволяют применять новые подходы к расширению запросов и документов. Однако большинство подходов тестируются на данных для английского языка, оцениваются на англоязычных бенчмарках, качество работы на разных типах русскоязычных текстовых данных неизвестно.

Данный проект посвящен созданию бенчмарка, т.е. совокупности датасетов для тестирования моделей информационного поиска для русского языка, а также тестированию моделей на созданном бенчмарке.

На текущий момент созданный бенчмарк содержит 27 датасетов различного происхождения, включая переведенные датасеты, датасеты из существующих мультязычных наборов данных, существующие русскоязычные датасеты, а также набор на основе созданных на основе интересных фактов Википедии.

В текущем году была опубликована в материалах конференции Диалог-2025 статья

Building russian benchmark for evaluation of information retrieval models / G. Kovalev, M. Tikhomirov, E. Kozhevnikov et al. // In Proceedings of the International Conference "Dialogue-2025. — 2025. (индексируется в Скопусе)

По результатам участия в конференции DAMDID-2025 две статьи рекомендованы к публикации в журналах.

ПРИЛОЖЕНИЕ А
Объем финансирования темы в 2025 году
Таблица А.1

Источник финанси- рования	Объем (руб.)	
	Получено	Освоено собственными силами
грант РНФ	1 500 000,00000	1 500 000,00000