ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему Влияние разрешения кореференции на задачу извлечения фактов из текстов разной сферы функционирования

Тема на английском The influence of coreference resolution on fact extraction from the texts of different domains

Студентка 3 курса
группы БКЛ212

Гришанова Анна Юрьевна
(Ф.И.О.)

Научный руководитель
Толдова Светлана Юрьевна
(Ф.И.О)
доцент
(должность, звание)

Москва, 2024 г.

Оглавление

1. Введение

Вопросно-ответные системы (Question Answering systems) – это такой тип поисковых систем, который как принимает вопросы, так и отвечает на них на естественном языке. На данный момент самой распространенной является двухступенчатая модель извлечения/считывания фактов (retriever/reader).

Несмотря на то, что в QA системы чаще всего разрабатываются для кратких ответов на вопросы, задача получения развернутых ответов так же важна: не на все вопросы можно ответить односложно. Развернутые ответы на вопросы можно получить, например, в результате решения задачи многодокументного реферирования, ориентированного на запросы (Query-Focused Multi-Document Summarization, QF-MDS). QF-MDS – это задача автоматического создания сводки (summary) из набора документов, которая отвечает на запрос конкретного пользователя. Сочетая в себе две задачи – извлечения фактов и составления связного текста – QF-MDS имеет большой прикладной потенциал. Кроме того, существенным преимуществом такого метода является то, что он обучается без учителя.

Метод QF-MDS является экстрактивным; он основан на выявлении, отборе и ранжировании предложений в соответствии с их соответствием заданному запросу – поэтому представление предложений и запросов является важным элементом QF-MDS-системы, влияющим на ее эффективность [Lamsiyah et al. 2021]. В работе [Lamsiyah et al. 2023] авторы доказывают: как выявление релевантных запросу пользователя предложений, так и генерация связных сводок более эффективны после разрешения анафоры.

Несмотря на то, что в дискурсивной теории анафора и кореференция являются разными явлениями, в компьютерной лингвистике задача разрешения анафоры близко связана с задачей разрешения кореференции. Более того, авторы датасета соревнования Ru-EVAL-2019 считают анафору одним из типов кореференции [Budnikov et al. 2019]. Авторы датасета приводят следующие типы кореференции:

1. Анафора

   (1) *To cook a turkey, you should first rub it all over with the butter*

2. Кореферентные именные группы

(2) *Some of our colleagues are going to be supportive. These kinds of people will earn our gratitude.*

3. Кореферентные глагольные группы

(3) *It is hard to change a human nature. It requires a lot of will.*

4. Раздробленные (split) антецеденты

(4*) *Bob and John are good friends. They go to school together.*

[Lamsiyah et al. 2023] в своей работе фокусируются исключительно на разрешении прономинальной анафоры (тип 1). Мы, в свою очередь, не останавливаемся на разрешении анафоры и затрагиваем также кореферентные именные группы. Здесь и далее *кореференцией* мы называем типы 1-2.

К нашему сведению, тема QF-MDS еще не было поднята в русскоязычной литературе; более того, мы не нашли русскоязычных датасетов, которые были бы полезны для решения этой задачи. Цель данной работы – проверить, улучшит ли разрешение кореференции качество работы системы QF-MDS на русском материале. Для достижения этой цели мы поставили следующие задачи:

1. создать русскоязычный датасет для решения задачи QF-MDS
2. проверить на русскоязычных данных введенный [Lamsiyah et al. 2021] метод QF-MDS
3. сравнить эффективность разрешения кореференции для решения задачи QF-MDS

2. Методы генерации ответов в вопросно-ответных системах: существующие решения и проблемы

Задача многодокументного реферирования, ориентированного на запросы, комбинирует в себе две задачи: ответ на вопрос (Question Answering, QA) и автоматическую суммаризацию. Рассмотрим существующие методы решения этих задач.

2.1. Существующие вопросно-ответные системы

Относительно простой способ получить ответ на вопрос – задать его большой языковой модели. Подав ей на вход строку с вопросом (префикс), мы можем попросить модель выполнить условную генерацию с учетом этого

префикса и считать вывод модели ответом. Такой метод часто применяют для ответов на простые фактивные вопросы; он опирается на то, что огромные предварительно обученные языковые модели во время обучения считывают множество фактов из обучающей выборки и кодируют эту информацию в своих параметрах.

У данного метода несколько недостатков. Во-первых, большие языковые модели *галлюцинируют*. *Галлюцинация* в этом смысле – ответ, который не соответствует реальных фактам. Галлюцинации появляются из-за того, что LLM-модели, получив на вход вопрос, просто придумывают ответ, который звучит разумно, а не опираются на какие-либо факты или источники. Так, исследуя, как LLM-модели отвечают на вопросы юридической тематики (например, о деталях отдельных дел), [Dahl et al. 2024] выяснили, что галлюцинации юридической тематики в больших языковых моделях преобладают над верными ответами: 69% в модели ChatGPT 3.5 и 88% в Llama 2.

Вторым недостатком применения LLM моделей для вопросно-ответных систем является частота их обновления. Вопросно-ответные системы особенно полезны для применения к относительно небольшим данным. Примером такого применения является ответ системы на вопрос пользователя по содержимому почтового ящика. Небольшие данные динамичны и быстро меняются, в то время как на обновление LLM модели могут уходить месяцы.

Принимая во внимание эти проблемы, наиболее популярным современным решением для построения вопросно-ответных систем является двухступенчатая модель извлечения/считывания (retriever/reader): на первом этапе из коллекции текстов извлекаются документы, в которых скорее всего содержится полезная для ответа на вопрос информация. Для следующего шага используются два разных метода: экстрактивный (extractive), при котором ответ извлекается из текстов документов, либо абстрактивный (abstractive), при котором используются LLM-модели для генерации ответа [Jurafsky, Martin 2024]. В настоящей работе мы применяем экстрактивный метод извлечения информации.

2.2. Автоматическая суммаризация

Суммаризация подразделяется на однодокументную (single-) и многодокументную (multidocument), а методы ее получения разделяются на

экстрактивные и абстрактивные. Остановимся на экстрактивной суммаризации. Этот метод подразумевает, что резюме (сводка, summary) полностью состоит из извлеченных ранжированных по релевантности ключевой информации данных. Экстрактивная суммаризация может быть выполнена как с помощью правил (примером является метод RIPTIDES для суммаризации новостей [White et al. 2001]), так и с помощью нейронных сетей (например, с помощью Feed Forward Neural Network [Bebis, Georgiopoulos 1994]).

Хорошая сводка должна соответствовать следующим критериям:

1. содержание ключевой информации
2. краткость
3. отсутствие избыточности
4. актуальность
5. связность (когезия) и читабельность [Verma et al., 2019].

Согласно теории дискурса, существуют следующие средства когезии: анафорические местоимения, лексические повторы, коннекторы (союзы) [Halliday, Hasan, Ruqayia 1976]. Экстрактивные системы, однако, структуру дискурса понимают плохо, и поэтому вследствие анафоры часть информации теряется. С другой стороны, отсутствие средств когезии в полученной сводке снижает связность и читабельность текста. Таким образом, в настоящей работе мы предполагаем, что разрешение кореференции улучшит соответствие сводок критериям 1 и 5.

2. Создание русскоязычного датасета

Для оценки качества QF-MDS систем для соревнований DUC 2005-2007 были составлены особые датасеты: [Dang 2005], [Dang 2006], [Dang 2007]. Каждый из этих датасетов устроен следующим образом. Составителями были выбраны 45-50 тем, далее для каждой темы было скачано по 25-50 релевантных статей из новостных источников. После этого эксперт формулировал развернутый вопрос на каждую из тем и добавлял бинарный параметр детализации ответа на него (*сжато* (*general*)*, развернуто* (*specific*)). Разметчики, опираясь на документы из созданной коллекции и ориентируясь на параметр детализации, составляли сводку-ответ на заданный вопрос; длина сводки не превышала 250 слов.

Подробная информация о количестве тем, размере коллекции документов, их источниках и числе написанных разметчиками сводок в каждом из датасетов представлена в Таблице 1. Пример темы из датасета DUC 2005 представлен в (5).

<table>
<tr><th>датасет</th><th>число тем</th><th>число документов на тему</th><th>источник документов</th><th>число сводок на тему</th><th>максимальный размер сводки</th></tr>
<tr><td rowspan="2">DUC 2005</td><td>20</td><td rowspan="2">25-50</td><td rowspan="2">Los Angeles Times, Financial Times of London</td><td>9</td><td rowspan="3">250 слов</td></tr>
<tr><td>30</td><td>4</td></tr>
<tr><td>DUC 2006</td><td>50</td><td rowspan="2">25</td><td rowspan="2">Associated Press, New York Times, Xinhua newswire</td><td rowspan="2">4</td></tr>
<tr><td>DUC 2007</td><td>45</td></tr>
</table>

Таблица 1. Подробная информация о содержании датасетов DUC 2005-2007.

(5) [Dang 2005]

тема*: American Tobacco Companies Overseas*

вопрос*: In the early 1990's, American tobacco companies tried to expand their business overseas. What did these companies do or try to do and where? How did their parent companies fare?*

детализация*: specific*

К нашему сведению, русскоязычного набора данных для решения задачи QF-MDS нет; перед нами стояла задача его создать. Как упоминают создатели датасетов DUC 2005-2007, мануально решать задачу QF-MDS сложно: на создание каждой сводки вручную у разметчиков уходило около 5 часов [Dang 2005]. Так, мы приняли решение создать датасет из доступных в сети Интернет развернутых ответов на вопросы.

Наше внимание привлекли вопросы после глав в школьных учебниках, а именно – в учебнике истории 5 класса [Вигасин и др. 2023]. У такого типа данных несколько преимуществ. Во-первых, разнообразие доступных ответов: их предоставляют как различные сайты с готовыми решениями задач из учебников, так и форумы, на которых пользователям отвечают другие пользователи (например, otvet.mail.ru). Во-вторых, заранее известная коллекция документов,

которые нужны для ответов на вопросы – это тексты глав учебника. Третье преимущество затрагивает именно учебники классов средней школы: вопросы в них предполагают скорее нахождение ответа в тексте, а не порождение его путем рассуждений. Получается, что такие данные хорошо подходят для нашей задачи экстрактивного многодокументного реферирования, ориентированного на запросы.

Мы отобрали 25 вопросов, классифицированных учебником [Вигасин и др. 2023] как “Проверь себя”, и собрали для каждого из них по 4 развернутых ответа. Тематическое разнообразие, аналогичное датасетам DUC, было сохранено: мы ограничивались выбором 1 вопроса из каждой главы.

Благодаря тому, что учебная программа является стандартизированной, содержание учебников разного авторства одинаково. Так, мы собрали тексты восьми разных учебников. Для того, чтобы еще более разнообразить коллекцию документов, была использована Википедия[1]: мы вводили в поиск название главы (например, *Афинская демократия при Перикле*) и скачивали все статьи, удовлетворяющие поиску. После этого мы мануально перепроверяли, есть ли в статье информация, нужная для ответа на вопрос.

Итак, созданный нами датасет устроен следующим образом. Коллекция документов состоит из 8 учебников по 50-60 глав и 62 документов из Википедии. В коллекции документов больше 40000 предложений, что сопоставимо с размерами коллекции документов датасетов DUC 2005-2007 [Dang 2007]. Всего в нашем датасете 25 вопросов-тем, и для каждого предложено по 4 “Золотые” сводки. Датасет доступен на странице GitHub[2].

3. Метод автоматической суммаризации

3.1. Описание метода

В настоящей работе мы тестируем эффективность метода QR-MDS, предложенного в [Lamsiyah et al. 2021]. Данный метод является unsupervised (не требует обучения с учителем); он основан на получении векторных представлений (эмбеддингов) предложений из документов и ранжирования их по

[1] https://ru.wikipedia.org/wiki

[2] ttps://github.com/ayugrishanova/The-influence-of-coreference-resolution-on-fact-extraction/

степени сходства с векторным представлением вопроса *Q*. Для ранжирования [Lamsiyah et al. 2021] применяют метод BM25 [Robertson et al. 1995], расчет косинусной близости и метод Максимальной Предельной Релевантности (Maximum Marginal Relevance method, MMR) [Carbonell, Goldstein 1998].

Согласно результатам, полученным в [Lamsiyah et al. 2021], эффективность данного метода превосходит многие аналогичные системы и показывает результаты, сравнимые с методами, основанных на глубоком обучении с учителем.

3.2. Получение векторных представлений

Высокая эффективность метода QR-MDS заключается в использовании качественной модели, порождающей векторные представления документов и запросов. Мы использовали модель SentenceBert (sBert) [Reimers, Gurevych 2019], которая является модификацией классической модели BERT [Devlin et al. 2018]. Несмотря на то, что BERT имеет state-of-the-art показатели на многих задачах регрессии для пар предложений (например, определение семантического сходства), решения таких задач требуют большой вычислительной мощности: авторы утверждают, что поиск наиболее похожей пары в корпусе из 10 000 предложений требует от BERT около 50 миллионов вычислений (~65 часов) [Reimers, Gurevych 2019].

До появления sBERT для решения этой проблемы исследователи индивидуально отображали каждое предложение точкой на семантическом пространстве, однако такой метод порождал недостаточно качественные эмбеддинги предложений. sBERT, в свою очередь, использует структуры сиамской и триплетной сетей, благодаря чему и происходит экономия вычислений. Пример структуры sBERT приведен на Рис. 1. Трансформер BERT принимает оба входа (предложения) параллельно, поэтому выходы не зависят друг от друга. Такая модификация позволяет этой архитуктуре не только быстро порождать эмбеддинги предожений без потери качества по сравнению с BERT, но и решать задачи, ранее неприменимые к BERT: крупномасштабное сравнение семантического сходства, кластеризация и извлечение информации с помощью семантического поиска.

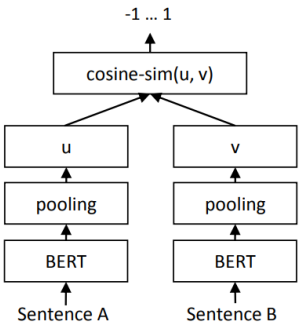

Рис. 1. Структура sBERT с целевой функцией регрессии (источник: [Reimers, Gurevych 2019]).

Применять sBERT можно для задач классификации и регрессии. Задача классификации заключается в выборе для каждой пары предложений логической связи между ними: *entailment* (логическое следствие), *contradiction* (противоречие) и *neutral* (см. пример (6)).

(6) предложение *Я – лягушка* имеет следующие логические связи с предложениями (6а-в):

а. *Я умею квакать* (логическое следствие)

б. *Я зеленая* (neutral)

в. *Я умею летать* (противоречие)

Задача регрессии заключается в предсказании правильной оценки семантической близости пары предложений; вычислять оценку можно с помощью косинусоидальной близости, Евклидового расстояния или скалярного произведения (dot-product).

Нашей задачей является поиск наиболее семантически близких вопросу *Q* предложений в документах, поэтому мы решаем с помощью sBERT задачу регрессии. Рассмотрим структуру, представленную на Рис. 1, подробнее. После передачи предложения в BERT к полученным эмбеддингам применяется mean-pooling слой для уменьшения размерности: исходные 512 768-мерных векторов преобразуются в один 768-мерный вектор. После получения таким образом векторов *u* и *v* прогнозируемая оценка сходства (в нашем случае – косинусоидальное сходство) сравнивается с истинным значением, и модель обновляется с помощью функции потерь MSE (средняя квадратичная ошибка).

Проект sBERT предоставляет различные модификации sBERT, предобученные для определенных задач. В нашей работе мы использовали модель

all-mpnet-base-v2[3]. Мы остановились именно на ней, потому что, согласно создателям проекта, она показывает лучшие результаты в задаче порождения эмбеддингов предложений[4].

Для оценки качества работы модели мы использовали датасет *nli-rus-translated-v2021* [Дале 2021]. Этот датасет составлен из различных англоязычных NLI (Natural Language Inference, Интерференция в Естественном Языке) датасетов, автоматически переведенных на русский язык. По нашим сведениям, аналогичных датасетов, созданных из “исконно” русскоязычных данных, еще не существует.

Предобученная модель *all-mpnet-base-v2* показывает не самые лучшие результаты: на тестовых данных датасета *nli-rus-translated-v2021* она предсказывает со средней квадратичной ошибкой 0,335. Мы обучали модель со следующими (гипер)параметрами:

(7) число эпох: {1, 3, 5} (см. Таблицу 2);

размер батча: 16;

warmup_steps: 50;

evaluation_steps: 250;

размер train датасета: 10000;

размер валидационного датасета: 500,

размер тестового датасета: 1000.

число эпох обучения	MSE
0 (модель “из коробки”)	0.335
1	0,171
3	0,103
5	0,041
7	0,046

Таблица 2. Результаты обучения модели в зависимости от числа эпох

Качество модели после обучения на 5 эпохах нас устроило, и далее в работе мы используем именно ее.

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[4] https://sbert.net/docs/pretrained_models.html

3.3. Ранжирование предложений.

После предварительной токенизации и лемматизации всех предложений в коллекции документов мы с помощью sBERT получили их векторные представления. Для ранжирования полученных эмбеддингов мы следовали методике, описанной в [Lamsiyah et al. 2021] и [Lamsiyah et al. 2023]. Согласно ей, ранжирование разделяется на следующие шаги:

1. Оценка схожести предложения с запросом и выбор *k* наиболее похожих на запрос предложений
2. Повторная оценка предложений для ликвидации повторов (избыточности)

Подробно рассмотрим эти шаги.

3.4.1. Первичная оценка предложений

[Lamsiyah et al. 2021] предлагают использовать следующую формулу для оценки релевантности каждого предложения для запроса:

(8) $score_{relevance}(S_i) = \alpha \times RSV_{bm25}(S_i, Q) + (1 - \alpha) \times RSV_{Sim}(S_i, Q)$, где

S_i – рассматриваемое предложение,

α – коэффициент значимости аргумента,

RSV_{bm25} (S_i, Q) – оценка релевантности S_i для Q согласно алгоритму BM25 [Robertson et al. 1995],

Q – запрос пользователя,

RSV_{Sim} – косинусное расстояние между S_i и Q.

BM25 считается одной из самых популярных вероятностных моделей поиска информации. Данный алгоритм основан на предположении о бинарной независимости, при которой вес элемента запроса (терма) вычисляется с использованием как частоты использования элемента как в предложении, так и в запросе. Стандартный BM25 имеет недостаток, который может быть релевантен для нашего исследования: в нем компонента нормализации частоты использования терма по длине документа не имеет надлежащего нижнего порога; в результате этого недостатка длинные документы, которые соответствуют термину запроса, часто могут быть оценены так же, как короткие документы, которые вообще не релевантны для запроса. Поэтому, в отличие от [Lamsiyah et al. 2021], мы используем модификацию BM+ [Trotman et al. 2014], которая решает эту проблему.

Комбинируя алгоритм BM25+ [Trotman et al. 2014] с оценкой косинусного расстояния, формула (8) учитывает как лексические, так и семантические признаки предложения и запроса. Коэффициент α дает возможность регулировать вклад того или иного типа признаков в оценку схожести. Авторы [Lamsiyah et al. 2021] утверждают, что $\alpha = 0.9$ показывает наилучшие результаты, поэтому в наших экспериментах мы использует именно это значение.

Для дальнейшей работы мы отбираем $top\_k = 50$ предложений с наивысшими оценками по формуле (8).

3.4.2. Повторная оценка для ликвидации повторов

Задачей QF-MDS является ответ на вопрос в виде сводки, поэтому ранжирование и выбор наиболее подходящего предложения недостаточен: для составления сводки важна не только релевантность предложения запросу, но и его новизна. Для придания выводу модели этих характеристик [Lamsiyah et al. 2021] используют метод Максимальной Предельной Релевантности (Maximum Marginal Relevance method, MMR) [Carbonell, Goldstein 1998]. Формально он определен формулой (9). Оценка MMR наиболее высокая, когда предложение имеет наибольшую схожесть с запросом пользователя и наименьшую схожесть с предложениями, уже выбранными для сводки.

(9)

$$score_{MMR}(S_i) = \mathrm{Argmax}_{S_i \in \text{top-}k \backslash Sum}\left[\lambda RSV_{Sim}(S_i, Q) - (1-\lambda)\max_{S_j \in Sum} CosSim(\vec{S_i}, \vec{S_j})\right]$$
$$1 \le i \le k, i \ne j$$

, где S_i – рассматриваемое предложение,

RSV_{Sim} – косинусное расстояние между S_i и Q,

Q – запрос пользователя,

top-k – группа из k наиболее близких запросу Q предложений,

Sum – группа уже выбранных для сводки предложений,

$top - k \backslash Sum$ – группа еще не выбранных предложений в *top-k*,

λ – интерполяционный коэффициент в диапазоне [0,1] с шагом 0.1.

Эффективность метода MMR особенно наглядна при сравнении полученных сводок до и после его применения (см. Таблицу 3): в левом столбце есть избыточные повторы информации.

ВОЗНИКНОВЕНИЕ ЗЕМЛЕДЕЛИЯ И СКОТОВОДСТВА. Почему сородичи слушались старейшин?	
до MMR	после MMR
Появление земледелия и скотоводства изменило жизнь людей. Возникновение республики. Но при этом уже было неважно, сородичи они между собой или нет. — возникновение города-государства Спарта. — возникновение города-государства Спарта. Почему греки покидали родину…	*Появление земледелия и скотоводства изменило жизнь людей. Возникновение республики. Но при этом уже было неважно, сородичи они между собой или нет. Гипотеза предполагает также привязку к этой общности возникновение письменности, укреплённых городов, разделения труда, централизованного управления, возникновение социальных классов и возникновение первой цивилизации на базе культуры Винча…*

Таблица 3. Сравнение работы алгоритма ранжирования до и после MMR

3.5. Составление сводок

После ранжирования предложений мы объединяем их, опираясь на несколько факторов. Во-первых, размер полученной сводки должен быть меньше 250 слов и не больше 150% средней длины ответа из Золотого Стандарта. Во-вторых, мы используем алгоритм сохранения хронологического порядка: предложения, встречающиеся в одном документе, в сводке стоят в порядке своего появления в нем. В-третьих, в сводке мы отражаем только предложения, состоящие больше, чем из 5 слов: таким образом мы элиминируем назывные предложения, не несущие достаточно информации (например, названия глав).

3.6. Оценка метода

Несмотря на хорошие показатели качества порождения векторных представлений, результаты работы модели оставляют желать лучшего: метрики R-1, R-2, R-SU показывают результаты 0.156, 0.024 и 0.119 соответственно. Мы решили проанализировать, какие ошибки совершает система. Посмотрим на пример сгенерированной сводки:

(10) запрос: *Индия и Китай в древности, природа и люди древней индии. Что представляли собой джунгли на берегах Ганга? Какую пользу они приносили людям? Какие опасности подстерегали там человека?*

1. *Всё должно подчиняться воле фараона — не только люди, но и сама природа.*

2. *Мужчины иногда приносили с охоты детёнышей убитых животных.*
3. *Её покрывали густые, труднопроходимые леса — джунгли.*
4. *В Древней Индии войско состояло из пехоты, конницы, колесниц и слонов.*
5. *Но в джунглях человека со всех сторон подстерегали опасности — хищные звери, ядовитые насекомые и змеи.*
6. *Они приносили царю дары от самых далёких племён.*

Во-первых, встречаются предложения, не соответствующие теме запроса (предложения 1, 2, 4, 6). Во-вторых, в предложениях встречаются анафорические элементы, антецедентов к которым нет в предшествующем контексте (подчеркнуты в предложениях 3, 6). В настоящей работе мы фокусируемся на решении проблемы неразрешенной кореференции.

3.7. Задача разрешения кореференции

Цель нашей работы подразумевает особую важность выбора модели для разрешения кореференции. Проводя предварительные эксперименты, мы пользовались моделью miniLM [Wang et al. 2020], дообученной для задачи разрешения конференции на датасете OntoNotes [Hovy et al. 2006]. Преимущество miniLM заключается в ее мультилингвальности: ее можно использовать без дообучения на русскоязычных данных. Метрики качества разрешения кореференции моделью miniLM, подсчитанные на тестовом датасете соревнования Ru-eval-2019 [Budnikov et al. 2019], представлены во второй строке Таблицы 4. Качество результатов работы модели miniLM заметно хуже результатов, полученных лучшей моделью соревнования Ru-eval-2019 [Budnikov et al. 2019].

модель	muc			bcube			ceafe		
	recall	precision	f-measure	recall	precision	f-measure	recall	precision	f-measure
best-Ru-eval-2019	82.62			73.95			72.14		
miniLM	41.83	67.54	51.66	34.52	63.34	44.69	41.77	48.93	45.07
RuCo-	51.98	70.91	59.99	44.89	64.08	52.80	48.58	58.36	53.02

BERT									

Таблица 4. Сравнение метрик качества моделей разрешения кореференции.

Так, чтобы добиться лучшего качества разрешения кореференции, мы использовали обученную на корпусе RuCoCo [Dobrovolskii et al. 2022] модель RuCo-BERT [Skobinsky 2022]. Эта модель показывает хорошие результаты на тестовой выборке корпуса RuCoCo [Dobrovolskii et al. 2022]: 81.1 precision, 78.2 recall и 79.6 F-1 score.

Модель RuCo-BERT ненамного лучше, чем miniLM, справляется с задачей разрешения кореференции на тестовой выборке корпуса Ru-eval-2019 (см. Таблицу 4). Ее главным недостатком было время работы: на обработку одного файла у модели уходило больше 1 минуты. Таким образом, мы остановились на miniLM как на оптимальном варианте для экспериментов.

Последним шагом в этапе разрешения кореференции метода QR-MDS является получение копий документов датасета, где в каждом предложении S_i члены кореферентной цепочки заменены своим главным антецедентом – первым упоминанием в кореферентной цепочке. Такой выбор главного антецедента имеет свои недостатки. Во-первых, первый член цепочки может иметь слишком большую длину, что ухудшает читабельность текста. Во-вторых, элементы кореферентной цепочки могут иметь разные грамматические характеристики (например, падеж). Оба этих недостатка продемонстрированы на примере (11):

(11) *Особенно велико значение археологических раскопок для изучения прошлого* ***людей, живших на земле до появления письменности****. Почти всё, что известно об* ***людей, живших на Земле до появления письменности****, рассказали находки археологов.* (текст с разрешенной кореферентностью)

Такие ошибки способны ухудшить как качество ранжирования документов, так и читабельность сводок. Повторы больших кластеров текста благодаря MMR понижаются в ранге, поэтому релевантное предложение с неудачно разрешенной кореференцией, как в (11), может быть ранжировано ниже нерелевантного.

Для того, чтобы решить данную проблему, мы составляем коллекцию документов как из текстов с разрешенной кореференцией, так и оригинальных текстов; выброс повторов из генерируемой сводки предусмотрен MMR.

4. Метрики качества

Следуя методике [Lamsiyah et al. 2023], мы оцениваем качество полученных сводок как количественно (ROUGE-метрики), так и качественно (человеческая оценка).

4.1. Количественная оценка

Для оценки автоматической суммаризации чаще всего используют метрики из семейства ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin 2003]. Такие метрики считают количество совпадений (overlaps) слов и словосочетаний в сгенерированном и тестовом резюме. ROUGE-метрики особенно хороши тем, что по большей части вычисляют recall (часть релевантной информации, сохраненная в сводке); это помогает убедиться в том, что важные детали из оригинального текста не были утеряны в процессе суммаризации.

Существуют разные типы ROUGE; следуя за [Lamsiyah et al. 2023], мы вычисляем:

1. ROUGE-1 (совпадения униграмм)
2. ROUGE-2 (совпадения биграмм)
3. ROUGE-SU (совпадения skip-грамм и униграмм).

4.2. Опрос носителей русского языка

Для post-оценки составленных сводок были привлечены 3 асессора; им предлагалось оценить по 4 варианта сводки для 4 вопросов. Вопросы были выбраны случайным образом.

Для каждого вопроса в случайном порядке предлагались следующие типы сводок:

1. Текст “Золотого” стандарта;
2. Текст, сгенерированный с помощью baseline на документах без разрешенной кореференции;
3. Текст, сгенерированный с помощью QR-MDS на документах без разрешенной кореференции;
4. Текст, сгенерированный с помощью QR-MDS на документах с разрешенной кореференцией.

Данные типы были выбраны в соответствии с задачами нашего исследования: сравнивая 2 и 3, мы проверяем эффективность метода QR-MDS; сравнивая 3 и 4,

мы проверяем влияние разрешения кореференции на релевантность и связность сводки.

Асессорам предлагалось оценить по шкале от 1 до 3 каждую из сводок в соответствии с ее полнотой, релевантностью представленной информации и связностью текста; таким образом мы проверяли соответствие сводок критериям, приведенным в **Разделе 2.2**:

1. содержание ключевой информации
2. краткость
3. отсутствие избыточности
4. актуальность
5. связность (когезия) и читабельность [Verma et al., 2019].

Одна из сводок, которые оценивали асессоры, представлена в **Разделе 9**.

5. Эксперименты

Мы оцениваем качество метода QF-MSD, сравнивая его с baseline-моделью. В качестве нее мы выбрали алгоритм BM25+ [Trotman et al. 2014], о котором писали в **Разделе 3**. После ранжирования BM25+ предложения проходили через те же шаги, которые были описаны в **Разделе 3.4**: первичная и повторная оценки релевантности. Векторные представления без использования sBERT в baseline были получены с помощью TF-IDF векторизации.

Для того, чтобы оценить влияние разрешения кореференции на прилагаемый метод, мы запускаем его на 3 вариантах коллекции документов:

1. Коллекция без разрешения кореференции.
2. Коллекция текстов с разрешенной кореференцией.
3. Сдвоенная коллекция как оригинальных текстов, так и текстов с разрешенной кореференцией.

Сравнивая качество метода на вариантах 2-3, мы можем оценить, как сильно проблема выбора неудачного главного антецедента (см. **Раздел 3.7**) влияет на эффективность суммаризации. Кроме того, мы отдельно запускаем метод QF-MSD на документах-учебниках и документах из Википедии. Так мы проверим, влияет ли размер коллекции документов на качество суммаризации.

6. Результаты и обсуждение

Результаты работы методов представлены в Таблице 5. Суффикс *+coref* означает, что коллекция документов состояла только из текстов с разрешенной кореференцией, суффикс *+coref+orig* означает, что коллекция документов была составлена как из текстов с разрешенной кореференцией, так и из оригинальных текстов.

	документы: учебник			документы: Википедия			полная коллекция документов		
	R-1	R-2	R-SU	R-1	R-2	R-SU	R-1	R-2	R-SU
baseline	0.123	0.010	0.094	0.105	0.002	0.083	0.116	0.005	0.104
baseline+coref	0.134	0.016	0.104	0.105	**0.004**	0.077	0.134	0.017	0.105
baseline+coref+orig	0.134	0.016	0.104	0.109	**0.004**	**0.084**	0.133	0.015	0.103
QF-MSD	0.125	0.012	0.096	**0.110**	0.001	**0.084**	**0.156**	**0.024**	**0.119**
QF-MSD+coref	**0.136**	**0.020**	**0.106**	0.104	**0.004**	0.078	0.145	0.015	0.109
QF-MSD+coref+orig	**0.136**	0.019	**0.106**	0.109	**0.004**	**0.084**	0.135	0.018	0.105

Таблица 5. Результаты работы моделей: baseline (BM+, ранжирование, MMR), QF-MSD (sBERT, ранжирование, MMR). Жирным выделены лучшие показатели метрики для каждой из коллекций документов.

Обсудим полученные результаты. Во-первых, средние результатов метрик ROUGE у моделей, запущенных на коллекции документов из Википедии меньше, чем у моделей, запущенных на коллекции документов-учебников. Согласно одностороннему парному t-тесту, это различие статистически значимо ($p < 0.001$). Так, мы можем сделать вывод, что большая коллекция документов приводит к лучшему качеству суммаризации.

Во-вторых, стоит заменить, что модель QF-MSD показывает лучшие результаты, чем baseline, однако статистическую значимость эта разница имеет только на коллекции документов-учебников (односторонний парный t-тест, $p < 0.005$). Кроме того, только на этой коллекции документов также видно заметное улучшение качества моделей (и baseline, и QF-MSD) после разрешения кореференции (односторонний парный t-тест, $p < 0.05$).

Итак, нельзя уверенно заявить ни что метод QF-MSD улучшает качество сводок, ни что разрешение кореференции повышает эффективность любой модели автоматической суммаризации.

Рассмотрим результаты post-оценки (Таблицы 6-8). Post-оценка проверяла, в первую очередь, соответствие резюме критериям хорошей сводки по [Verma et al. 2019].

	сводка 1	сводка 2	сводка 3	сводка 4	*средн. знач.*
gold	**2,67**	2,33	**2,67**	**3,00**	*2,67*
baseline	1,67	**2,67**	2,00	1,67	*2,00*
QF-MSD	2,33	**2,67**	2,33	1,33	*2,17*
QF-MSD+coref	1,67	2,33	2,00	1,67	*1,92*
средн. знач.	*2,08*	*2,50*	*2,25*	*1,92*	

Таблица 6. Оценка асессорами полноты сводки. gold – “золотой стандарт”. Жирным выделены лучшие результаты в каждой из сводок.

	сводка 1	сводка 2	сводка 3	сводка 4	*средн. знач.*
gold	**2,67**	**2,67**	2,33	**2,33**	*2,50*
baseline	2,33	2,00	**2,67**	1,67	*2,17*
QF-MSD	2,33	2,33	**2,67**	2,00	*2,33*
QF-MSD+coref	2,33	2,00	2,00	1,33	*1,92*
средн. знач.	*2,42*	*2,25*	*2,42*	*1,83*	

Таблица 7. Оценка асессорами релевантности сводки. gold – “золотой стандарт”

	сводка 1	сводка 2	сводка 3	сводка 4	*средн. знач.*
gold	2,67	**3,00**	**3,0**0	**3,00**	*2,92*
baseline	**3,00**	**3,00**	2,00	2,00	*2,50*
QF-MSD	2,67	2,67	2,33	2,00	*2,42*
QF-MSD+coref	1,67	1,33	2,33	1,67	*1,75*
средн. знач.	*2,50*	*2,50*	*2,42*	*2,17*	

Таблица 7. Оценка асессорами связности и читабельности сводки. gold – “золотой стандарт”

Интересно, что “Золотой стандарт” сводки не всегда получает самую высокую оценку. Так, для сводки 2 лучшую полноту показали полученные baseline и QF-MSD результаты; самую высокую релевантность для сводки 3

показали полученные baseline и QF-MSD результаты; для сводки 1 лучшую читабельность показал результат baseline. Мы считаем, что такие результаты получились случайно. Например, в сводке 1 в нескольких предложениях есть союзы-коннекторы (*при этом, но*). Наличие этих средств когезии никак не могло повлиять на ранжирование предложений: стоп-слова из предложений были удалены.

Для релевантности подтверждается отрицательное влияние разрешения кореференции: после применения одностороннего парного t-теста выборки полноты и связности получили оценку $p < 0.05$. Влияние QF-MSD метода на полноту ответа, его релевантность и связность не подтвердилась односторонним парным t-тестом.

Проблема неудачно выбранного главного антецедента (см. **Раздел 3.7**), превалирующая в сводках, основанных на документах с разрешенной кореференцией, снижает качество полученных резюме. Авторы [Lamsiyah et al. 2023] получают противоположный результат: генерация сводок с разрешением анафоры показывает статистически значимое улучшение по сравнению с генерацией финального текста без разрешения анафоры. Разница результатов может объясняться методикой разрешения кореференции: в нашей работе мы заранее готовим тексты с разрешенной кореференцией и применяем к ним метод QF-MSD. В свою очередь, [Lamsiyah et al. 2023] применяют pipeline-структуру. Таким образом они имеют возможность для каждого из предложений сохранить оригинальное местоимение в том случае, если в предложениях перед ним уже есть его антецедент.

7. Выводы

Итак, в настоящей работе мы предложили датасет для задачи многодокументного реферирования, ориентированного на запросы. На этом датасете мы проверили эффективность метода решения задачи QF-MDS, предложенного в [Lamsiyah et al. 2021]. Оказалось, что практически всегда метод QF-MDS и его модификации показывают лучшие значения согласно метрикам семейства ROUGE [Lin 2004]. Тем не менее, разница в качестве генерации сводок по сравнению с baseline-моделью не является статистически значимой.

Мы выяснили, что разрешение кореференции отрицательно влияет на релевантность информации в сводках (значимая разница по сравнению со сводками без разрешенной кореференции, $p < 0.05$). Такой эффект появляется из-за проблемы неудачного выбора главного антецедента кореферентной цепочки. Мы считаем, что решение данной проблемы – интересная тема для дальнейших исследований влияния разрешения кореференции на задачу многодокументного реферирования, ориентированного на запросы.

8. Список литературы

Вигасин и др. 2023 – А. А. Вигасин, Г. И. Годер, И. С. Свенцицкая. *История. Всеобщая история. История древнего мира: 5 класс: учебник*. М.: Просвещение, 2023.

Дале 2021 – Д. Дале. Нейросети для Natural Language Inference (NLI): логические умозаключения на русском языке. (электронный документ) https://huggingface.co/datasets/cointegrated/nli-rus-translated-v2021

Skobinsky 2022 – E. Skobinsky. RuCo-BERT: Coreference resolution model for Russian fully compatible with AllenNLP package. available at: https://github.com/gleb-skobinsky/RuCo-BERT.

Budnikov et al. 2019 – A. E. Budnikov et al. Ru-eval-2019: Evaluating anaphora and coreference resolution for russian. *Computational Linguistics and Intellectual Technologies-Supplementary Volume*. 2019.

Bebis, Georgiopoulos 1994 – G. Bebis, M. Georgiopoulos. Feed-forward neural networks. *Ieee Potentials* 13.4, 1994. P. 27-31.

Carbonell, Goldstein 1998 – J. Carbonell, J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries // Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, August 24-28 1998.

Dahl et al. 2024 – M. Dahl et al. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*. 2024.

Devlin et al. 2018 – Devlin J. et al. *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805. 2018.

Dobrovolskii et al. 2022 – V. Dobrovolskii, M. Michurina, and A. Ivoylova. RuCoCo: a new Russian corpus with coreference annotation. *arXiv preprint arXiv:2206.04925*. 2022.

Halliday, Hasan 1976 – M. Halliday, R. Hasan. *Cohesion in English*. London: Longman, 1976.

Honnibal, Montani 2017 – M. Honnibal, I. Montani. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. 2017.

Hovy et al. 2006 – E. Hovy et al. OntoNotes: the 90% solution // Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers. 2006.

Jurafsky, Martin 2024 – D. Jurafsky, J. Martin. *Speech and Language Processing (3rd ed. draft)*. available at: https://web.stanford.edu/~jurafsky/slp3/.

Lamsiyah et al. 2021 – S. Lamsiyah et al. Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, BM25 model, and maximal marginal relevance criterion. *Journal of Ambient Intelligence and Humanized Computing*, 2023. P. 1-18.

Lamsiyah et al. 2023 – S. Lamsiyah et al. Can Anaphora Resolution Improve Extractive Query-Focused Multi-Document Summarization? *IEEE Access*, 2023.

Lin 2004 – Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries // Text summarization branches out, Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, July 25 - 26, 2004. P. 74–81.

Reimers, Gurevych 2019 – Reimers N., Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv:1908.10084v1*, 2019.

Robertson et al. 1995 – S. Robertson et al. Okapi at TREC-3. *Nist Special Publication Sp* 109, 1995. P. 109.

Wang et al. 2020 – W. Wang et al. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems* 33, 2020. P. 5776-5788.

White et. al 2001 – M. White et al. Multidocument summarization via information extraction // Proceedings of the first international conference on Human language technology research, San Diego, California, USA, March 18-21, 2001.

9. Приложение

Вопрос 1. *ВОЗВЫШЕНИЕ АФИН В V ВЕКЕ ДО Н.Э. И РАСЦВЕТ ДЕМОКРАТИИ. Кто считался афинским гражданином? Кто — переселенцем? Кто — рабом? Что вы узнали о положении афинских граждан, переселенцев, рабов?*

Золотой стандарт: *Гражданином по закону считался коренной афинянин, у которого оба родителя (отец и мать) были гражданами. Граждане имели право участвовать в управлении государством. Переселенцами же считались люди из других городов, жившие в Афинах. Они говорили по-гречески, одевались так же, как и остальные афиняне, поклонялись тем же богам, но в отличие от граждан должны были платить пошлину за право проживания в Афинском государстве. Рабы были чужеземцами, взятыми в плен или купленными за морем, или захваченными пиратами. Рабы не имели прав человека: они не распоряжались своим трудом, не могли иметь семью, менять место жительства, даже имя раба зачастую менялось по воле хозяина.*

Результат baseline: *— В I веке до н.э., - сказал Источниковед, - Римская держава столкнулась с сильными противниками. Большинство афинских граждан были грамотными людьми. Расцвет древнегреческих полисов приходится на V век до н. э. При этом, во второй половине V века до н.э. В V веке до н.э., после победоносного завершения греко-персидских войн, архитектор Либон возвел здесь величественный храм Зевса Громовержца — Зевса Олимпийского. В V (5) — IV (4) веках до н.э. — законы вавилонского царя Хаммурапи, защищающие свободных граждан от превращения в рабов. Но вот в XIV (14) веке до н.э. Всех граждан Афин разделили на разряды в соответствии с их доходом. Периэки — свободные не-граждане (примерный аналог афинских метеков) Наивысший расцвет афинской демократии наступил в конце греко-персидских войн. Он защищал интересы большинства афинских граждан и внёс*

большой вклад в развитие демократии. Он провел закон, по которому гражданином считался только тот, у кого отец и мать были афинянами.

Результат QF-MSD: *Он защищал интересы большинства афинских граждан и внёс большой вклад в развитие демократии. — законы вавилонского царя Хаммурапи, защищающие свободных граждан от превращения в рабов. А гражданином считался мужчина, достигший 20 лет, у которого и мать, и отец были афинянами. Периэки — свободные не-граждане (примерный аналог афинских метеков) Большинство афинских граждан были грамотными людьми. После Грекоперсидских войн начинается расцвет Афин. После подавления мятежа у союзников отбирали земли и селили на них афинских граждан. Перикл ужесточил доступ в состав афинских граждан. Всех граждан Афин разделили на разряды в соответствии с их доходом. — В I веке до н.э., - сказал Источниковед, - Римская держава столкнулась с сильными противниками. Но вот в XIV (14) веке до н.э. В V (5) — IV (4) веках до н.э. В V веке до н.э., после победоносного завершения греко-персидских войн, архитектор Либон возвел здесь величественный храм Зевса Громовержца — Зевса Олимпийского.*

Результат QF-MSD и разрешения кореференции: *Демос — народ Афин — получил власть (по-гречески — «кратос») в Демос — народ Афин — городе, поэтому такой способ управления назвали демократией. Большинство афинских граждан были грамотными людьми. А гражданином считался мужчина, достигший 20 лет, у которого и мать, и отец были афинянами. Двадцать тысяч афинских рабов перебежали к врагу. Всех граждан Афинах разделили на разряды в соответствии с Всех граждан Афин доходом. — законы вавилонского царя Хаммурапи, защищающие свободных граждан от превращения в рабов. Но вот в XIV (14) веке до н.э. Перикл защищал интересы большинства афинских граждан и внёс большой вклад в развитие демократии. Наивысший расцвет афинской демократии наступил в конце греко-персидских войн. Среди купцов было немало переселенцев из других городов, постоянно проживавших в Афин. — В I веке до н.э., - сказал Источниковед, - Римская держава столкнулась с сильными противниками. В V (5-м) веке до н. э. главным портом Афин стал Пирей, расположенный в 5-6 км от Афин.*