

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»

Шикунова Ксения Максимовна

“Распознавание и генерация объектов сравнения в вопросно-ответных системах”
“Object Recognition and Generation for Comparative Question Answering Systems”

Выпускная квалификационная работа студентки 4 курса

Академический руководитель
образовательной программы
канд. филологических наук, доц.

Ю. А. Ландер

Научный руководитель
доцент Школы лингвистики

Э. С. Клышинский

« » _____ 2023 г.

Москва
2023

Оглавление

1. Введение	1
2. Обзор существующих подходов.....	3
2.1. <i>Comparative Argumentative Machine.....</i>	<i>3</i>
2.2. <i>Генерация ответов на сравнительные вопросы с использованием системы SAM</i>	<i>4</i>
3. Данные.....	6
3.1. <i>Набор данных для оценки работы моделей на этапе распознавания объектов и аспекта сравнения в вопросительных предложениях</i>	<i>6</i>
3.2. <i>Данные для подачи в SAM</i>	<i>7</i>
3.3. <i>Данные для оценки задачи суммаризации выдачи SAM</i>	<i>7</i>
4. Распознавания объектов и аспекта сравнения с использованием генеративных моделей	9
4.1. <i>Методология.....</i>	<i>9</i>
4.2. <i>Оценка работы моделей</i>	<i>10</i>
5. Генерация ответа на сравнительные вопросы с использованием дообученной модели GPT-2.....	18
5.1. <i>Методология.....</i>	<i>18</i>
5.2. <i>Оценка работы моделей</i>	<i>21</i>
6. Анализ результатов и возможные улучшения	26
6.1. <i>Оценка распознавания объектов и аспекта сравнения</i>	<i>26</i>
6.2. <i>Оценка генерации суммаризации</i>	<i>26</i>
6.3. <i>Возможные улучшения.....</i>	<i>28</i>
7. Заключение.....	28
Приложения.....	30
Литература	32

1. Введение

В настоящее время существует множество успешных вопросно-ответных систем, базирующихся на поиске данных, соответствующих предмету поиска, в закрытых корпусах или в веб-документах. Тем не менее, существует потребность в ответах на более сложные вопросы, например, вопросы мнения или сравнительные вопросы, которые составляют чуть меньше 5% запросов¹, получаемых наиболее популярными поисковыми системами. Такие вопросы в свою очередь тоже требуют сбора фактических данных о предмете сравнения, а также нуждаются в анализе мнений пользователей, у которых есть опыт работы с данным объектом. Из такого набора данных о предмете потенциально можно получить аргументированный ответ в пользу того или иного объекта сравнения.

Вопросные-ответные системы на данный момент делятся на три типа: они могут быть основаны на алгоритме поиска ответа по конкретному тексту, искать ответ по базе знаний (KBQA или Knowledge Base Question Answering, где источником знаний является база знаний, содержащая в себе набор данных в определенном формате), а также на основе поиска по большому объему текстового материала в виде множества документов с последующим объединением найденной информации в один текстовый ответ (ODQA или Open Domain Question Answering). К сожалению, эти вопросно-ответные системы не предполагают решение задачи сравнения, в частности такую конфигурацию данных о характеристиках или опыте использования объектов, требуемых для сравнения пользователями. По этой причине вопрос создания успешной вопросно-ответной системы (CQAS или Comparative Question Answering System) интересует многих исследователей. Уже существуют сравнительные системы, которые способны находить фактическую информацию об объектах и выдавать ее пользователю на самостоятельную оценку (такие как Diffzi², Diffen³ или Versus⁴). Большинство таких систем ограничены конкретной предметной областью, например, позволяют сравнивать технику по техническим характеристикам или бренды в общих чертах. Тем не менее такие системы не предоставляют пользователю исчерпывающий ответ на заданный вопрос и требуют от него самостоятельной итоговой оценки. Более

¹ Bondarenko et al., 2020

² <https://diffzi.com/>

³ <https://www.diffen.com/>

⁴ <https://versus.com/>

того, такие системы не могут удовлетворить более узкие вопросы в тех случаях, когда пользователю необходимо сравнить объекты в определенном аспекте их характеристик или использования.

Безусловно, в настоящее время уже были произведены попытки создать такую сравнительную систему, которая успешно могла бы имитировать человека-эксперта в широком спектре областей. Такая система должна не только осуществлять быстрый поиск информации об объектах по большим корпусам данных, но и уметь находить сравнение искомых предметов, а также корректно обрабатывать запрос и выдавать ответ на естественном языке. Первый шаг в сторону создания такой сравнительной вопросно-ответной системы был предпринят Chekalina et al. (EACL 2021), ими была создана система на основе Comparative Argumentative Machine (CAM⁵), в которой были реализованы как идентификация объектов и аспектов сравнения во входных свободных вопросах, так и несколько подходов для генерации ответа.

Инструмент CAM принимает на вход эксплицитно указанные объекты сравнения в виде строк, а также отдельно указанный аспект сравнения, если такой необходим, на выходе же выдает список найденных сравнительных высказываний об искомых объектах, разделенный на наборы в пользу одного объекта и в пользу другого. Инструмент, созданный Chekalina et al. (EACL 2021), для распознавания объектов и аспекта в запросе берет за основу токенизацию и лейблинг отдельных токенов, будь то объект, аспект или предикат, наша же система основана на генеративных моделях-трансформерах (Vaswani et al. 2017), которые получают на вход инструкцию по распознаванию нужных токенов в конкретном вопросе. Для генерации ответа на естественном языке в обоих исследованиях также используются разные стратегии, подробнее они описываются в главах Обзор существующих подходов и Методология.

Структура нашей работы будет состоять из определения цели работы и задачи исследования, разработки методологии и описания данных, а также оценки результатов и определения направления будущей работы.

Таким образом целью нашего исследования является изучить предложенную Chekalina et al. (EACL 2021) систему, выявить плюсы и минусы ее работы, а также предпринять попытку в сторону улучшения полученных в ходе первичного исследования результатов. Нашей же задачей является разработать вопросно-ответную систему для английского языка с применением нейросетевых моделей, которая

⁵ Schildwächter et al., 2019

позволит упростить использование САМ за счет внедрения механизма понимания вопроса на естественном языке и распознавания в нем объектов и аспекта сравнения с помощью генеративной модели, а также генерации выдачи суммаризации полученных с помощью инструмента САМ данных об искомых объектах.

Первым этапом нашей работы будет распознавание объектов и аспекта сравнения в сравнительных предложениях на английском языке, для этого мы оценим работу нескольких генеративных моделей-трансформеров и выберем ту, которая покажет наилучшие результаты. Следующим этапом будет подача найденных объектов и аспекта сравнения в САМ, после чего на основе результатов работы САМ мы будем генерировать осмысленные предложения на естественном языке в виде суммаризации первых двадцати предложений выдачи с реферированием к конкретным высказываниям. Последний этап нашей работы будет состоять в представлении демонстрационной версии системы. Используемые данные и код будет опубликован на GitHub.

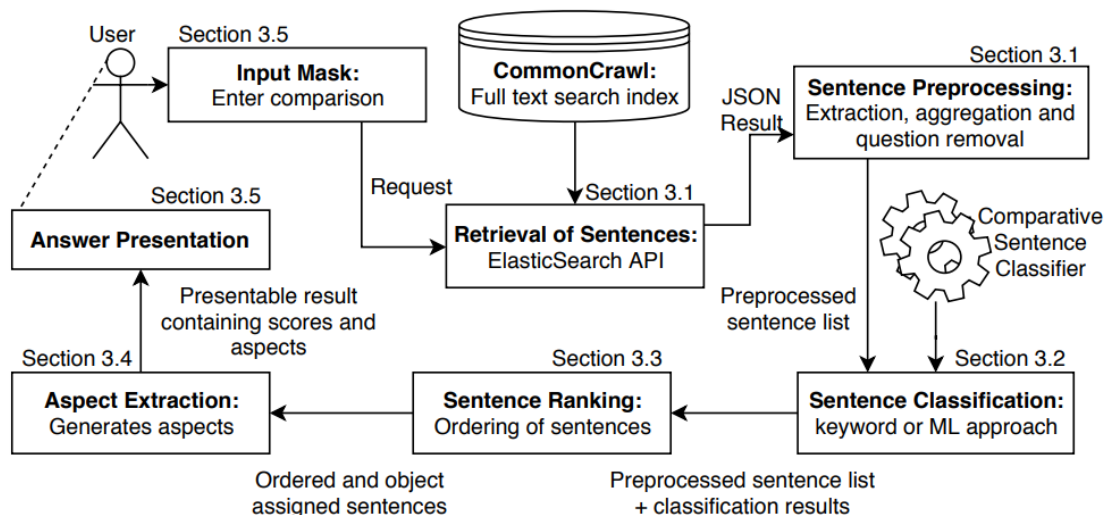
2. Обзор существующих подходов

2.1. Comparative Argumentative Machine

Comparative Argumentative Machine является системой поиска информации по большому массиву данных с целью аргументативного сравнения искомых объектов исходя из информации, извлеченной из веб-архива Common Crawl⁶.

Схема данной системы выглядит следующим образом:

Рисунок 1. Алгоритм работы Comparative Argumentative Machine



⁶ <https://commoncrawl.org/>

В первую очередь происходит поиск предложений, подходящих объектам из опроса, затем из найденного списка извлекаются предложения, содержащие сравнение обоих объектов, то есть предложения с текстовыми конструкциями, которые могли бы утверждать, что один из сравниваемых объектов превосходит другой, или что объекты равны или же что их нельзя сопоставить. Далее найденные высказывания ранжируются и из них извлекаются аспекты сравнения, после чего пользователю предоставляется ответ в виде списков высказываний в пользу каждого объекта.

Поиск предложений осуществляется на предварительно обработанном корпусе DepCC (Panchenko et al. 2017), который содержит 14,3 миллиарда английских лингвистически размеченных предложений. Предложения извлекаются за счет полнотекстового индекса Elasticsearch BM25⁷. Вопросительные высказывания в ходе поиска игнорируются, поскольку не могут влиять на получение аргументированного ответа. Далее отбираются предложения, в которых упоминаются оба объекта сравнения, после чего найденное предложение соотносится с одним из четырех классов, соответствующих найденной информации: один из объектов хуже, лучше, равен другому или сравнение не найдено.

Затем отсортированные предложения ранжируются исходя из оценки Elasticsearch и достоверности классификатора, ранг высказывания также увеличивается и в том случае, если аспекту сравнения пользователем присвоен высокий коэффициент. Более того САМ дополнительно генерирует до десяти аспектов, которые присваиваются объекту с более высокой частотой совпадения.

На выходе пользователь получает общее представление о найденной информации в виде шкал баллов по каждому аспекту, а также список высказываний «за» и «против» каждого аспекта сравнения.

2.2. Генерация ответов на сравнительные вопросы с использованием системы САМ

Система, созданная в статье Chekalina et al. (EACL 2021), разработана с той же целью, что и наш проект, и представляет собой расширенную версию САМ с помощью NLU (Natural Language Understanding) и NLG (Natural Language Generation) модулей. Цель же состоит в том, чтобы создать такой инструмент, который способен предоставить пользователю аргументированный ответ на вопрос сравнения, в полной мере описав при этом преимущества и недостатки каждого из сопоставляемых

⁷ <https://www.elastic.co/>

объектов. Данная система разделена на три этапа: (1) выделение объектов сравнения, аспекта и предиката из предложения на естественном языке; (2) передача найденных данных в САМ и получение результатов в виде высказываний «за» и «против» каждого объекта; и (3) генерация связных человеческих ответов.

Первый этап является попыткой автоматического распознавания нужной для использования САМ информации в предложениях на естественном языке и сводится к задаче sequence labeling. Для обучения моделей авторами был создан датасет Comparely (Chekalina et al. 2021), включающий в себя чуть более 3000 предложений, размеченных вручную. Предложения содержали в себе данные из трех областей, а именно области брендов, общих объектов и области компьютерных наук. Было протестировано 5 моделей, в число которых вошли однослойный двунаправленный LSTM (Arora et al. 2017); BiLSTM, расширенный CRF (Conditional Random Field (Sutton, McCallum 2012)); LSTM с ELMO (Peters et al. 2018); BERT (Devlin et al. 2019) и RoBERTa (Liu et al. 2019), которая показала наилучший результат по F1 в определении объектов, аспектов и предикатов (0.925, 0.685, 0.894 соответственно), в связи с чем была выбрана для внедрения в систему.

Последний же этап был представлен в виде двух подходов: генерация ответа с помощью предобученных генеративных моделей Transformers (а именно CTRL (Keskar et al. 2019)) и поиск информации с помощью ранжирования (САМ или TextRank (Mallick et al. 2019), контекстного дополнения (контекст из 3 предложений из документа, откуда взято первое высказывание в пользу «победителя») или построения шаблонов.

Для оценки работы генеративного модуля был выбран датасет L6–Yahoo! Answers Comprehensive Questions and Answers version 1.0⁸ (1200 отобранных сравнительных предложений) и метрики ROUGE (Lin 2004). К сожалению, мы не смогли получить доступ к данному набору данных, поэтому его использование в нашем исследовании невозможно. Несмотря на то, что модель CTRL получила самый высокий балл ROUGE-1, то есть выдавала наиболее близкие высказывания к текстам «лучших ответов» из датасета Yahoo!, ее ответы нельзя считать достаточно информативными для пользователей и структурированными, так как модель могла не предоставлять конкретного итогового выбора между объектами и никак не учитывала выдачу САМ, в которой как раз таки и содержалась нужная информация.

⁸ <https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11>

В результате для демонстрационного интерфейса были выбраны три подхода для генерации, которые получили лучшие оценки по ROUGE-N, между которыми пользователь мог переключаться самостоятельно, а именно ранжирование CAM, генерация с помощью CTRL и контекстное дополнение из документа первого высказывания из выдачи CAM.

3. Данные

3.1. Набор данных для оценки работы моделей на этапе распознавания объектов и аспекта сравнения в вопросительных предложениях

На первом этапе перед нами стояла задача оценки выдачи генеративных моделей, а именно оценки качества распознавания объектов и аспектов сравнения в сравнительных вопросах. Для этого нами был выбран набор предложений из набора данных Dataset-CompQA⁹ (Beloucif et al. 2022), в который вошло чуть более 3000 вручную размеченных предложений. В каждом предложении, представленном в датасете, каждому слову соответствует один из пяти тэгов: OBJ-1, OBJ-2, что соответствуют первому и второму объекту соответственно, ASPECT, что соответствует слову, выражающему аспект сравнения, SHARED – тэг, говорящий о том, что слово является частью первого и второго объектов сравнения, хоть и упоминается один раз, и O – означающий, что слово или знак не является ни аспектом, ни объектом сравнения. Приведем несколько примеров, иллюстрирующих расстановку тэгов в предложении (слова и знаки, не находящиеся в квадратных скобках, имеют тэг O):

(1) *who is [better ASPECT] [Manny OBJ-1] or [Ortiz OBJ-2] ?*

(2) *How can you tell the difference between [literal OBJ-1] and [symbolic OBJ-2] [dreams SHARED] ?*

Предложения в наборе данных были распределены следующим образом:

Таблица 1. Статистика набора данных Dataset-CompQA

	Аспект		Общее слово в объектах	
	выражен	не выражен	есть	нет
Количество предложений	1783	1274	727	2330
Процентное соотношение	58.3	41.7	23.8	76.2

⁹ <https://github.com/uhh-lt/Dataset-CompQA/>

Для удобства использования набора данных, мы преобразовали данные из tsv-формата (со столбцами, соответствующими слову или знаку и его тэгу) в json-формат, где каждый элемент содержит инструкцию для генеративной модели – текст предложения с добавлением в начало самой инструкции, то есть вопроса «What are the objects and the aspect of comparison in the sentence ‘{вопрос}’», список с объектами сравнения, содержащимися в вопросе, и аспектом. Так же для удобства оценки сгенерированных ответов мы преобразовали тот же набор данных в csv-формат со столбцами, соответствующими тексту предложения, первому объекту, второму объекту и аспекту сравнения.

3.2. Данные для подачи в САМ

На втором этапе данные представляли собой список двух объектов сравнения и найденный аспект, если такой существует, и подавались напрямую в САМ¹⁰, выдача которого представляла собой список из 10 сравнительных высказываний «за» и «против» для каждого из объектов сравнения.

3.3. Данные для оценки задачи суммаризации выдачи САМ

На третьем же этапе нам вновь необходимо было создать нужный формат данных для подачи инструкции генеративным моделям (более подробное описание которых содержится в следующем разделе). Нами был создан датасет в jsonlines-формате на основе выдачи Chat-GPT (Floridi, Chiriatti 2020) по запросу суммаризации двадцати приведенных аргументов в пользу каждого объекта сравнения, каждый элемент в котором представлял из себя набор из двух объектов сравнения, ввода для модели и трех примеров ожидаемого вывода. Набор данных имел следующие характеристики:

Таблица 2. Статистика набора данных на основе выдачи модели GPT-3

Средняя длина ввода	Средняя длина выдачи	Количество примеров ввода	Количество примеров выдачи
2456	1057	536	1608

Вводные данные включали в себя инструкцию и список предложений из выдачи САМ, например, для объектов «georgia» и «virginia» информация выглядела следующим образом:

¹⁰ <https://uhh-lt.github.io/cam/>

(3) *Write a comparison of "georgia" and "virginia". Summarize only relevant arguments from the list.*

1. *And Georgia is probably better than Virginia Tech.*

2. *Hopefully, in Georgia, he has a better staff than he did in Virginia.*

...

20. *Virginia, which likely has a better defense than UM, gave up 56 on Saturday to Georgia Tech.*

After the summary, list the arguments you used below the text. Put citations in brackets inside the text. Do not even mention arguments that are not relevant to georgia and virginia.

Мы добавили несколько примеров ожидаемого вывода модели для каждой пары объектов. Таким образом, для пары объектов выше один из ожидаемых выводов выглядел следующим образом:

(4) *Georgia and Virginia are two states in the United States, with their own unique characteristics and qualities.*

Some arguments mention that Georgia is probably better than Virginia Tech [1], has a relatively freer political climate [5][9], and has a good football team [4]. Others argue that Virginia has a nicer climate [15], lighter accents [12], and a better defense [20].

Some comparisons are also made between different schools from the two states, such as Georgia Tech's offense being better than Virginia Tech's defense [7], and Tech being better than FSU and Georgia Tech [11].

Ultimately, it's hard to say which state is "better" as it depends on individual preferences and experiences. Both states have their own unique history, culture, and geography that makes them worth exploring.

Arguments used: 1, 4, 5, 7, 9, 11, 12, 15, 20.

Далее во избежание обучения генеративной модели на нелогичных или ошибочных высказываниях мы вручную разметили часть набора данных на наличие логических ошибок при реферировании к высказываниям. Для каждого предложения в выдаче мы разметили правильно указанные номера примеров (с помощью тэга True), ошибочно указанные (тэгом False) и не содержащие информации о сравнении (тэгом None). Затем из каждого предложения мы удалили референсы, помеченные тэгами None или False, следующим образом: (1) если генерируемое высказывание имеет хотя бы один верный референс, неверные ссылки удалялись без изменения предложения; (2) если предложение основывалось на единственной неверной ссылке, убирались и сам

референс, и часть высказывания, опирающаяся на него, или, в отдельных случаях, высказывание целиком.

4. Распознавания объектов и аспекта сравнения с использованием генеративных моделей

4.1.Методология

Для генерации ответа на запрос об объектах и аспекте сравнения в сравнительном вопросе мы выбрали семь моделей на основе архитектуры трансформера. Для шести моделей мы использовали два подхода подачи данных: zero-shot learning и few-shot learning. В первом случае модели задавался один вопрос с инструкцией, то есть, например:

(5) *What are the objects and the aspect of comparison in the sentence 'who is better manny or ortiz ?'*

Во втором же случае мы подавали на вход несколько предложений с примерами ответов в одинаковом формате и только после этого задавали свой вопрос. Для примеров мы выбрали такие предложения, в которых отображается разнообразие возможных структур вопросительных высказываний. Поскольку время обработки запроса моделью увеличивалось пропорционально количеству добавленных к вопросу примеров, мы остановились на четырех из них. Таким образом мы взяли следующие примеры: (1) с общим словом (с тэгом SHARED); (2) без общего слова, но с разной длиной объектов; (3) со сложно выраженным аспектом сравнения; (4) состоящий из нескольких предложений. Также к этим примерам были добавлены ответы в формате, который мы ожидаем от модели при выдаче ответа на вопрос. Таким образом пример запроса для сравнительного вопроса из примера (5) в few-shot формате выглядел так:

(6) *What are the objects and the aspect of comparison in the sentence 'How can you tell the difference between literal and symbolic dreams ?'*

Objects: 'literal dreams', 'symbolic dreams'

Aspect: -

What are the objects and the aspect of comparison in the sentence 'is a family med graduate as efficient as a medicine one ?'

Objects: 'family med graduate', 'medicine'

Aspect: efficient

What are the objects and the aspect of comparison in the sentence 'Wouldn't you rather have your kids and their friends drunk in your house than out on the street ?'

Objects: 'in your house', 'on the street'

Aspect: kids drunk

What are the objects and the aspect of comparison in the sentence 'Who do you think the U.S. Secret Service likes more ? Democrats or Republicans ? Why do you think ?'

Objects: 'Democrats', 'Republicans'

Aspect: likes

What are the objects and the aspect of comparison in the sentence 'who is better manny or ortiz ?'

Итак, мы попробовали обучить выбранные нами модели без дополнительного обучения и без изменения стандартных параметров. Следующим шагом была обработка выдачи каждой из моделей в соответствии с определенными паттернами генерации каждой из них, далее мы подробнее рассмотрим процесс обработки выдачи каждой из моделей.

4.2. Оценка работы моделей

После приведения результатов работы модели в удобный для оценки формат, а именно эксплицитное указание названных объектов и аспекта для каждого предложения вне сгенерированных моделью фраз, мы выбрали несколько стратегий оценки результатов генерации:

1) Полное совпадение: найденные объекты и аспект считались правильно распознанными только при полном совпадении строк с действительными объектами или аспектом сравнения из датасета;

2) Косинусная близость: на основе эмбедингов действительных и предсказанных объектов и аспекта сравнения, полученных с помощью библиотеки sentence-transformers (Reimers, Gurevych 2019) и модели bert-base-nli-mean-tokens (Reimers, Gurevych 2020), рассчитывалась средняя косинусная близость для объектов сравнения и аспекта;

3) Косинусная близость исключительно по найденным элементам;

4) Расстояние Левенштейна: среднее число перестановок, полученное с помощью инструмента `edit_distance` библиотеки `nltk`¹¹, требующихся выдачи модели для соответствия с действительными объектами и аспектом сравнения в тех случаях, когда предсказание модели верно;

5) Message Understanding Conference (MUC 1995);

6) Constrained Entity Alignment F-Measure (Luo 2005).

Стоит также отметить, что для получения более точной оценки работы моделей в выделении аспекта сравнения мы не учитывали те слова, которые не несли в себе специального лексического значения. Другими словами, мы считали, что в том случае, если аспект выражен сравнительным словом (*than*, *as* и проч.) или словом *better*, то он отсутствует.

Для оценки работы самих моделей мы использовали *relaxed*-метрики, тем не менее для точного определения, подходит ли модель для следующих этапов мы оценивали модель строже, так как, например, лишь одного найденного моделью объекта сравнения недостаточно для перехода на второй этап нашей системы, поскольку для корректного поиска по веб-корпусу с помощью САМ необходимо передать на вход оба объекта сравнения.

Таким образом мы получили следующие результаты для объектов сравнения:

Таблица 3. Оценки работы моделей в распознавании объектов сравнения: косинусная близость и полное совпадение

		Полное совпадение		Косинусная близость	
		Relaxed	Strict	Relaxed	Strict
t5-large	zero-shot	0.0	0.0	0.0	0.0
	few-shot	0.649	0.418	0.700	0.699
	few-shot+	0.503	0.356	0.250	0.246
flan-t5-large	zero-shot	0.321	0.205	0.656	0.595
	few-shot	0.437	0.382	0.679	0.559
gpt2	zero-shot	0.0	0.0	0.0	0.0
	few-shot	0.326	0.166	0.605	0.553
gpt-neo-1.3B	zero-shot	0.0	0.0	0.0	0.0
	few-shot	0.392	0.276	0.738	0.695
dolly-v2-3b	zero-shot	0.098	0.064	0.247	0.219

¹¹ <https://www.nltk.org/>

	few-shot	0.0	0.0	0.0	0.0
llama-7b-hf	zero-shot	0.358	0.299	0.619	0.617

Таблица 4. Оценки работы моделей в распознавании объектов сравнения: косинусная близость и расстояние Левенштейна, рассчитанные исключительно по найденным объектам

		Косинусная близость	Расстояние Левенштейна
t5-large	zero-shot	0.0	-
	few-shot	0.863	4.64
	few-shot+	0.793	3.875
flan-t5-large	zero-shot	0.757	8.92
	few-shot	0.827	4.4
gpt2	zero-shot	0.0	-
	few-shot	0.741	7.58
gpt-neo-1.3B	zero-shot	0.0	-
	few-shot	0.783	6.75
dolly-v2-3b	zero-shot	0.768	10.82
	few-shot	0.0	-
llama-7b-hf	zero-shot	0.852	7.65

Таблица 5. Оценки работы моделей в распознавании объектов сравнения: MUC и CEAF

		MUC			CEAF		
		precision	recall	f-score	precision	recall	f-score
dolly	zero-shot	0.239	0.264	0.222	0.195	0.199	0.181
	few-shot	0.0	0.0	0.0	0.0	0.0	0.0
neo-gpt	zero-shot	0.0	0.0	0.0	0.0	0.0	0.0
	few-shot	0.793	0.571	0.604	0.712	0.546	0.585
gpt-2	zero-shot	0.0	0.0	0.0	0.0	0.0	0.0
	few-shot	0.600	0.515	0.497	0.519	0.475	0.464
llama	zero-shot	0.584	0.598	0.549	0.480	0.479	0.455
t5	zero-shot	0.0	0.0	0.0	0.0	0.0	0.0
	few-shot	0.849	0.725	0.725	0.673	0.645	0.624

	few-shot+	0.781	0.769	0.712	0.449	0.673	0.499
flan-t5	zero-shot	0.663	0.751	0.616	0.556	0.568	0.517
	few-shot	0.764	0.791	0.724	0.636	0.591	0.592

А также следующие результаты для аспектов сравнения:

Таблица 6. Оценки работы моделей в распознавании аспекта сравнения: косинусная близость и полное совпадение

		Полное совпадение	Косинусная близость
t5-large	zero-shot	0.0	0.0
	few-shot	0.318	0.498
	few-shot+	0.166	0.469
flan-t5-xl	zero-shot	0.0	0.0
	few-shot	0.59	0.704
gpt2	zero-shot	0.0	0.0
	few-shot	0.200	0.397
gpt-neo-1.3B	zero-shot	0.0	0.0
	few-shot	0.274	0.496
dolly-v2-3b	zero-shot	0.425	0.477
	few-shot	0.0	0.0
llama-7b-hf	zero-shot	0.412	0.575

Таблица 7. Оценки работы моделей в распознавании аспекта сравнения: косинусная близость и расстояние Левенштейна, рассчитанные исключительно по найденным объектам

		Кос. близ. (не учит. не найд.)	Расстояние Левенштейна
t5-large	zero-shot	0.0	-
	few-shot	0.805	8.56
	few-shot+	0.779	10.34
flan-t5-xl	zero-shot	0.0	-
	few-shot	0.926	3.01
gpt2	zero-shot	0.0	-
	few-shot	0.769	9.44
gpt-neo-1.3B	zero-shot	0.0	-

	few-shot	0.835	7.63
dolly-v2-3b	zero-shot	0.936	8.01
	few-shot	0.0	-
llama-7b-hf	zero-shot	0.867	19.52

Таблица 8. Оценки работы моделей в распознавании аспекта сравнения: MUC и CEAF

		MUC			CEAF		
		precision	recall	f-score	precision	recall	f-score
dolly	zero-shot	0.444	0.471	0.445	0.241	0.264	0.243
	few-shot	0.0	0.0	0.0	0.0	0.0	0.0
neo-gpt	zero-shot	0.0	0.0	0.0	0.0	0.0	0.0
	few-shot	0.391	0.421	0.369	0.321	0.336	0.309
gpt-2	zero-shot	0.0	0.0	0.0	0.0	0.0	0.0
	few-shot	0.315	0.292	0.272	0.294	0.280	0.266
llama	zero-shot	0.477	0.561	0.475	0.325	0.388	0.327
t5	zero-shot	0.0	0.0	0.0	0.0	0.0	0.0
	few-shot	0.406	0.443	0.386	0.330	0.359	0.321
	few-shot+	0.363	0.415	0.351	0.299	0.346	0.295
flan-t5	zero-shot	0.0	0.0	0.0	0.0	0.0	0.0
	few-shot	0.683	0.624	0.629	0.477	0.431	0.440

Далее следует подробнее рассмотреть процесс обработки данных выдачи каждой оцениваемой нами модели.

Одной из испытываемых моделей мы выбрали Large Language Model Meta AI (LLaMA (Touvron et al. 2023)) – NLP модель с 13 миллиардами весов на основе архитектуры трансформера. К сожалению, в связи с нехваткой вычислительных мощностей нам удалось проверить ее работу только с помощью zero-shot подхода.

Как видно из таблиц выше, модель не оправдала наших ожиданий в полной мере. Тем не менее нам удалось выделить четыре стратегии генерации модели:

- Модель генерировала вопросительное предложения без инструкции;
- Модель повторяла весь запрос;
- Модель генерировала осмысленный ответ на запрос в формате «The objects are {objects} and the aspect is {aspect}»;

- Модель генерировала ответ с использованием команд LaTeX¹² для неупорядоченных списков, которые включали в себя строки «`\begin{itemize}`», «`\item`», «`\end{itemize}`» и «`\end{code}`».

Следующей моделью мы выбрали Generative Pre-trained Transformer 2 (или GPT-2 (Radford et al. 2019)) – модель на архитектуре трансформера с 1.5 миллиардами параметров. К сожалению, результаты выдачи модели при zero-shot подходе преимущественно оказались бессмысленными, модель или просто повторяла запрос, или генерировала несвязные высказывания на основе набора слов из запроса, например:

(7) *Q: What are the objects and the aspect of comparison in the sentence 'How can you make the work output of a machine GREATER than the work INPUT ?':*

A: The answer is that the work output of a machine GREATER than the work INPUT is the same as the work output of a machine.

С использованием few-shot подхода результаты улучшились. Выдача модели состояла из повторения запроса и ответа в формате «Objects: {objects} Aspect: {aspect}», тем не менее модель плохо справилась с самой задачей определения объектов и аспекта сравнения. Наиболее частотные ошибки, которые допускала модель, состоят в следующем:

- Модель добавляла контекстное окружение к выделенному аспекту;
- Модель находила только один объект;
- Модель находила более двух объектов или разделяла один объект на несколько частей.

Модель GPT-Neo (Black et al. 2021) – немного доработанная версия модели GPT-2, обученная на наборе данных The Pile (Gao et al. 2020), также не показала хорошего результата с zero-shot вариантом запроса. Модель скорее продолжала генерацию подобных вопросов, а не отвечала на ответ, в связи с чем мы изменили инструкцию и ушли от вопросительного формата в сторону повествовательного, требующего продолжения – новый вариант запроса начинался с фразы «In the sentence {предложение} there are two objects and the aspect of comparison:». Тем не менее такой вариант инструкции не улучшил выдачу.

¹² <https://www.latex-project.org/>

С few-shot подходом удалось получить результаты в ожидаемом формате, тем не менее выдача практически ничем не отличалась от результатов работы предыдущей модели.

Следующие две модели хуже всего справились с задачей распознавания объектов и аспекта сравнения в сравнительных вопросах.

GPT-J (Wang, Komatsuzaki 2021) – модель, похожая на GPT-3, обученная, как и GPT-Neo, на наборе данных The Pile, так же как и предыдущая модель отвечала довольно развернуто, но в результате обоих подходов не предоставляла необходимой нам информации, по причине чего в оценочной таблице информация о результатах ее работы отсутствует. Например, в случае zero-shot запроса, модель могла сгенерировать следующий ответ:

(8) *What are the objects and the aspect of comparison in the sentence 'How can you tell the difference between literal and symbolic dreams?'?*

I know that literal and symbolic dreams are the same subject and it should be "How can you tell the difference between the dreams and the symbols they represent". I'm also aware that when you use 'between' you are using the objects as the subject.

А в случае few-shot подхода повторяла запрос и начинала придумывать похожие примеры вопроса с вымышленным ответом [9] или выдавала ответ вне запрашиваемого формата [10]:

(9) *What are the objects and the aspect of comparison Trading of what?*

Objects: 'of what', 'of what'

Aspect: trading

(10) *The correct answer should really be "The sentence doesn't make sense".*

Dolly (Conover et al. 2023) – дообученная на Стэнфордском датасете версия модели GPT-J, справилась с задачей лучше своей предшественницы в результате zero-shot подхода, тем не менее модель предоставляла ответ в произвольном формате, а повторяющиеся конструкции предложений встречались крайне редко, поэтому парсинг полученных данных не оказался в полной мере успешным. Few-shot подход оказался еще менее продуктивным, так как модель часто генерировала ответы на вопросы из примеров.

Модели text-to-text transfer transformer, или T5 (Raffel et al. 2022) и Flan-T5 (Chung et al. 2022) (T5 с обучением Flan-PaLM), показали наилучшие результаты с

подходом few-shot. Модели придерживались нужного формата выдачи и вполне успешно находили объекты и аспекты сравнения. Также вторая модель неплохо показала себя в выделении объектов сравнения при zero-shot подходе, тем не менее аспект упоминался в высказывании без определенного паттерна, поэтому оценить работу модели по результатам выполнения этой задачи не представлялось возможным.

Мы предприняли несколько попыток улучшить результаты работы этих моделей. В первую очередь было решено увеличить количество примеров во few-shot подходе до девяти, добавив при этом примеры, где: (1) аспект вводится с помощью наиболее распространенной формулировки *in terms of*, (2) аспект выражен глаголом, (3) аспект не выражен, (4) сравнение выражено вопросом о качестве соотношения объектов сравнения, (5) объекты имеют общее слово, которое имеет различное местонахождение в словосочетании. Такие примеры были выбраны с учетом того, что аспекты в среднем определялись хуже объектов, а также по причине того, что в случае с общим словом оно довольно часто опускалось или присваивалось лишь одному из объектов.

Эксперимент был проведен на модели T5, в результате чего мы получили менее хорошие результаты, как видно из Таблиц 1 и 2 в графе few-shot+.

Далее мы выдвинули гипотезу, что модели имеют определенные паттерны ошибок, поэтому подсчитали процент неверно распознанных объектов и аспектов, разделив выборку на 4 категории: (1) наличие общего слова в объектах сравнения, (2) отсутствие общей части в объектах, (3) наличие аспекта сравнения, (4) отсутствие аспекта. Как мы выяснили, за исключением последнего пункта, модели допускают примерно одинаковое количество ошибок в остальных случаях, поэтому было решено дообучить обе модели, случайно разделив набор данных для оценки в соотношении 80 и 20 процентов для тренировочной и тестовой выборки соответственно. Далее мы оценили модели с помощью выбранных метрик:

Таблица 9. Оценка работы дообученных моделей T5 и Flan-T5: полное совпадение и косинусная близость

		Полное совпадение		Косинусная близость	
		Relaxed	Strict	Relaxed	Strict
t5	Objects	0.678	0.606	0.942	0.941
	Aspect	0.691		0.808	
flan-t5	Objects	0.689	0.609	0.942	0.942

	Aspect	0.688	0.808
--	--------	-------	--------------

Таблица 10. Оценка работы дообученных моделей T5 и Flan-T5: косинусная близость и расстояние Левенштейна, рассчитанные исключительно по найденным объектам

		Косинусная близость	Расстояние Левенштейна
t5	Objects	0.945	3.29
	Aspect	0.949	2.33
flan-t5	Objects	0.947	3.35
	Aspect	0.949	2.06

Таблица 11. Оценка работы дообученных моделей T5 и Flan-T5: MUC и CEAF

		MUC			CEAF		
		precision	recall	f-score	precision	recall	f-score
t5	Objects	0.914	0.922	0.888	0.708	0.716	0.694
	Aspect	0.776	0.773	0.749	0.529	0.528	0.516
flan-t5	Objects	0.913	0.932	0.895	0.706	0.719	0.697
	Aspect	0.776	0.779	0.753	0.528	0.529	0.517

Результаты распознавания объектов и аспекта сравнения в вопросительных предложениях заметно улучшились.

5. Генерация ответа на сравнительные вопросы с использованием дообученной модели GPT-2

5.1. Методология

Для суммаризации полученных высказываний «за» и «против» объектов сравнения в контексте конкретного аспекта, если такой существует, мы выбрали несколько версий предшествующей GPT-3 модели, а именно GPT-2, с учетом наших вычислительных возможностей, требуемых для обучения данной модели. Предобработанный нами набор данных, состоящий из 50 вручную размеченных текстов, мы использовали как обучающую выборку, мы также сгенерировали дополнительные 50 примеров ввода для тестовой выборки, чтобы оценить результаты работы модели.

Мы посчитали необходимым ручную разметку текстов, поскольку сгенерировавшая их модель допускала довольно частые ошибки в реферировании к

текстам, тем не менее всегда сохраняла необходимый формат выдачи. С помощью ряда экспериментов с дообучением предшествующих ChatGPT моделей мы планировали получить более осмысленные и логически верные высказывания с более успешным реферированием к примерам.

В ходе разметки мы оценили точность реферирования к списку аргументов моделью GPT-3, таким образом каждая верная ссылка на высказывание в пользу того или иного аргумента (то есть ссылка с тэгом True) считалась за единицу TP (True Positive), остальные ссылки, а именно реферирование с тэгами None и False были отнесены к FP (или False Positive) единицам. Таким образом точность реферирования моделью ChatGPT была оценена в 0.898, что не так уж и плохо, тем не менее не является достаточным для использования в качестве идеальной обучающей выборки.

На исправленных данных мы обучили две версии модели GPT-2 с 124 миллионами и с 345 миллионами параметров, примеры результатов обучения которых представлены ниже соответственно:

- (11) *Nokia and Siemens are both major technology companies with their respective strengths and weaknesses.*

Some argue that Nokia is better than Samsung in this segment [1]. Others argue that Samsung is better in terms of software development [4], and that Nokia has better engineering teams [7]. The choice of which company to prioritize in this segment depends on personal preferences and priorities.

Arguments used: 1, 2, 3, 4, 6, 7, 9, 12, 15, 18.

- (12) *Nokia and Siemens are both well-known manufacturers of electronic devices, and their performances are often compared.*

According to some arguments, Nokia is a better choice for certain tasks, such as mobile phone connectivity and location services, or integration of cloud services [1]. On the other hand, some argue that Siemens has better design and ergonomics [3][7][8], and is better in terms of integration of cloud services [17].

There are also general arguments about the quality of workmanship between the two companies, with some stating that Nokia is quicker to fix problems [12], and makes

better phones [7]. Additionally, some argue that Nokia has a better balance sheet [5][6], and has better results in certain types of examinations [9].

Overall, the choice between Nokia and Siemens depends on the specific needs and preferences of the user. It's important to do research and compare different models to get a better understanding of their performance and features before making a decision.

Arguments used: 1, 2, 3, 5, 6, 7, 8, 9, 12, 15, 17, 19.

Как мы видим, модели ссылаются к абсолютно разным примерам, тем не менее ни одна из моделей не справляется ни с задачей перечисления использованных примеров в конце своего ответа, ни с задачей логичного реферирования, так как большинство ссылок указывают на примеры, в которых содержится иная информация об объектах сравнения. Точность реферирования для обеих версий модели составила менее 0.4. Тем не менее стоит отметить, что мы не можем напрямую сравнивать точность реферирования GPT-3 и дообученных версий GPT-2, поскольку размеченные наборы данных в количественном соотношении сильно разнились, а также сам подход к генерации моделей отличался, другими словами, в первом случае модель производила генерацию с нуля, во втором же – мы обучали модель генерировать выдачу в конкретном, возможно, не типичном для нее формате, поэтому будем считать оценку 0.4 примерной, но не точной.

Для примера рассмотрим ссылки к высказываниям 1 и 7:

(13) *1. Some people say Nokia mobile is better than Siemens because it's easier to operate.*

(14) *7. So yes, Nokia is now falling faster than at its worst moment Palm, Motorola, Siemens, Windows Mobile, Ericsson, Blackberry or any other maker.*

В первом случае обе версии модели правильно указывают, что об объекте *Nokia* говорится в положительном ключе, тем не менее ни одна из них не предоставляет в предложении с реферированием ту же информацию, что содержится в примере, а именно аспект сравнения *easier to operate*.

Во втором же случае вторая версия модели дважды ссылается на седьмой пример, более того в одном варианте допускает ошибку как с определением оттенка высказывания, так и с описанием аспекта сравнения.

Тем не менее можно также заметить, что модель не игнорирует примеры и не генерирует совсем необоснованные выводы. Так, например, в примере (12) в третьем предложении говорится об эргономичности продуктов, что упоминалось в примере 1, поэтому с точки зрения задачи суммаризации без учета успешности реферирования модель работает успешно.

Стоит отметить, что главной целью данного этапа являлось как раз таки создание модуля, способного успешно суммаризировать полученную информацию, реферирование же напротив нашей главной целью не являлось, хотя такая функция могла бы быть довольно полезна и заметно бы упростила анализирование информации пользователем.

Мы намеренно решили не использовать в итоговом инструменте ChatGPT, поскольку данная модель на данный момент не находится в открытом доступе и к тому же слишком тяжела для частного использования. Поэтому нашим решением стало использование легковесной общедоступной модели GPT-2, которая наиболее близка к доработанной своей версии GPT-3. Более того мы обучили выбранную модель на данных с высоким показателем точности, а также на вручную обработанном сокращенном наборе данных.

5.2. Оценка работы моделей

Таким образом мы предприняли несколько подходов в обучении модели и выбрали метрики ROUGE и BLEU (Papineni et al. 2002) для сравнения результатов выдачи модели с золотым стандартом, который представлял из себя в разных экспериментах либо скорректированные данные выдачи ChatGPT, либо с необработанной генерацией этой модели. Целью автоматической оценки являлось понять, насколько сгенерированный текст по смыслу подобен проверенному, и являются ли результаты выдачи GPT-2 обоснованными, повторяющими информацию об объектах и аспекте сравнения из высказываний, найденных с помощью CAM.

Стоит также отметить, что данные метрики были выбраны как наиболее подходящие, но не идеальные, поскольку обучение модели GPT-2 проходило на формате данных выдачи, удобном для другой модели, в частности ChatGPT, и на основе полученных оценок нельзя обсуждать, работает ли выбранная для генерации

модель лучше или хуже своего потомка. То же самое можно сказать и о входных данных, так как инструкция тоже была подобрана под работу GPT-3.

В качестве первого эксперимента мы оценивали те же модели, для которых была подсчитана точность, то есть те, которые были обучены на 50 отредактированных примерах. Тестовый набор данных состоял из дополнительных 50 примеров, сгенерированных ChatGPT и не подвергавшихся разметки после. Ниже представлены оценки метрик ROUGE и BLEU для пар примеров суммаризации GPT-3 и разных дообученных версий GPT-2.

Таблица 12. Оценка ROUGE по результатам первого эксперимента

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
GPT-2 124M	0.437	0.145	0.266	0.383
GPT-2 345M	0.462	0.171	0.289	0.405

Таблица 13. Оценка BLEU по результатам первого эксперимента

	BLEU	precision				brevity penalty	length ratio
GPT-2 124M	0.196	0.51	0.25	0.14	0.08	1.0	1.013
GPT-2 345M	0.217	0.58	0.31	0.19	0.11	0.886	0.892

Как видно из представленных данных, оценка ROUGE на основе униграмм у обеих версий модели GPT-2 довольно высока, тем не менее остальные версии метрики указывают на недостаточно успешные результаты, а именно оценка на основе биграмм и на основе самой длинной общей последовательности символов в отдельных предложениях и полном тексте. Оценка BLEU в данном случае оказалась неудовлетворительной. Версия модели с бóльшим количеством параметров показывает чуть лучшие результаты, но ненамного.

Мы также предприняли попытку оценить работу модели, обученной на части размеченных данных, сравнительно с отредактированными примерами выдачи ChatGPT так же с помощью метрик ROUGE и BLEU. Для этого мы разделили размеченный набор данных в соотношении 4 к 1 на обучающую и тестовую выборки соответственно. На основе десяти тестовых примеров были получены следующие оценки:

Таблица 14. Оценка ROUGE по результатам второго эксперимента

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
GPT-2 124M	0.444	0.157	0.266	0.389
GPT-2 345M	0.469	0.182	0.304	0.411

Таблица 15. Оценка BLEU по результатам второго эксперимента

	BLEU	precision				brevity penalty	length ratio
GPT-2 124M	0.183	0.58	0.29	0.16	0.08	0.837	0.849
GPT-2 345M	0.196	0.62	0.33	0.19	0.10	0.778	0.799

Данный эксперимент также не дал достаточно хороших результатов, оценки хоть и немного выше, но практически не отличаются. К тому же модель, обученная на меньшем количестве данных, начала генерировать более короткие ответы.

Следующий эксперимент состоял в том, чтоб дополнительно, уже во второй раз, дообучить модели на тех же 50 примерах, но без предварительной обработки. То есть в итоге модель была обучена на 100 примерах, где каждые 50 различаются десятью процентами точности. Гипотеза состояла в том, чтобы сначала обучить модель предоставлять данные определенного формата, а затем обратить ее внимание на ошибки, которых хотелось бы избежать при генерации суммаризации. Тестовая выборка состояла из дополнительных автоматически сгенерированных размеченных 50 примеров с отличающимся содержанием. Результаты оценки соответствующих метрик представлены в виде таблиц ниже:

Таблица 16. Оценка ROUGE по результатам третьего эксперимента

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
GPT-2 124M	0.449	0.163	0.286	0.394
GPT-2 345M	0.471	0.182	0.299	0.415

Таблица 17. Оценка BLEU по результатам третьего эксперимента

	BLEU	precision				brevity penalty	length ratio
GPT-2 124M	0.224	0.54	0.29	0.17	0.10	0.967	0.968
GPT-2 345M	0.240	0.54	0.29	0.18	0.11	1.0	1.053

В ходе данного эксперимента обе версии модели вновь улучшили свои показатели, более того длина сгенерированной суммаризации практически совпала с эталонной.

Мы также предприняли стратегию аугментации данных за счет расширения набора данных при обучении. Обучающая выборка на этот раз состояла из 50 размеченных и скорректированных вручную примеров и дополнительно еще из 50 примеров автоматической генерации, имеющих другое содержание, в результате чего итоговая обучающая выборка, как и в предыдущем эксперименте, составила 100 примеров. Гипотезой в этом случае являлся тот факт, что модели для корректной генерации не хватало большего количества разнообразных примеров, поэтому мы прибегли к расширению набора уникальных примеров с учетом небольшого ухудшения качества данных в связи с дополнением в виде 50 нескорректированных примеров выдачи. Мы посчитали такое ухудшения качества данных незначительным, поскольку точность такой выдачи, как говорилось ранее, составляла около 0.9.

Таблица 18. Оценка ROUGE по результатам четвертого эксперимента

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
GPT-2 124M	0.449	0.161	0.279	0.398
GPT-2 345M	0.489	0.180	0.303	0.423

Таблица 19. Оценка BLEU по результатам четвертого эксперимента

	BLEU	precision				brevity penalty	length ratio
GPT-2 124M	0.202	0.58	0.31	0.18	0.10	0.829	0.842
GPT-2 345M	0.216	0.611	0.325	0.187	0.106	0.861	0.869

Расширение и разнообразие обучающего набора данных также привело к улучшению результатов, метрика ROUGE на основе униграмм показала оценку, близкую к 0.5, балл за геометрическую точность метрики BLEU оказался около 0.6 у обеих версий модели, тем не менее средняя длина выдачи не повторяет эталонную.

Последний эксперимент также исходил из идеи расширить обучающую выборку, поскольку данный подход показал наилучший результат. Для этого при обучении модели мы использовали полный автоматически сгенерированный и неразмеченный набор данных, за исключением первых 50 примеров, которые удалось исправить вручную. Таким образом обучающая выборка составила 450 примеров пар ввода и

сгенерированной ChatGPT суммаризации высказываний об объектах сравнения, а тестовая выборка составила 50 предобработанных примеров. Результаты обучения представлены также в таблице ниже:

Таблица 20. Оценка ROUGE по результатам пятого эксперимента

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
GPT-2 124M	0.495	0.202	0.318	0.429
GPT-2 345M	0.541	0.225	0.343	0.469

Таблица 21. Оценка BLEU по результатам пятого эксперимента

	BLEU	precision				brevity penalty	length ratio
GPT-2 124M	0.255	0.57	0.32	0.19	0.12	1.0	1.029
GPT-2 345M	0.289	0.61	0.36	0.23	0.14	1.0	1.002

В последнем эксперименте с наибольшим размером обучающей выборки нам наконец удалось получить удовлетворительные результаты, практически все версии метрики ROUGE имеют хорошие показатели, оценка BLEU для GPT-2 среднего размера практически дошла до 0.3, хотя геометрическая точность практически не изменилась. Более того обе версии модели генерируют текст, совпадающий по длине с эталонным, что может также говорить и об отсутствии сильного разброса в длине генерируемых текстов.

Стоит также подробнее обратить внимание на то, по какой причине мы считаем, что оценка BLEU со значением 0.3 считается хорошей. По данным [eff.org](https://www.eff.org/ai/metrics#Translation)¹³ в задаче, например, машинного перевода, в сравнении с профессиональным ручным переводом машинный перевод может достигать в среднем до 25-30 баллов по BLUE-score, когда сам ручной перевод оценивается в 50 баллов, подробный график можно найти в Приложении 1.

Что касается метрики ROUGE, то результат от 40 баллов также является уже достаточно хорошим по данным [nlpprogress.com](http://nlpprogress.com/english/summarization.html)¹⁴ из раздела с исследованиями на тему суммаризации текстов. Так, например, в статье BRIO (Liu et al. 2022) ROUGE-1 достигает 47.78 баллов, а ROUGE-2 и ROUGE-L 23.55 и 44.57 баллов соответственно.

¹³ <https://www.eff.org/ai/metrics#Translation>

¹⁴ <http://nlpprogress.com/english/summarization.html>

6. Анализ результатов и возможные улучшения

6.1. Оценка распознавания объектов и аспекта сравнения

Одной из наших задач являлось создать инструмент, позволяющий упростить использование системы САМ и предоставить пользователю возможность ввода сравнительного вопроса в более дружелюбном для него формате. Мы попробовали использовать различные генеративные модели на базе архитектуры трансформеров для решения задачи понимания вопроса на естественном языке и распознавания объектов и аспекта сравнения. В первую очередь мы оценили работу семи разных генеративных моделей с помощью двух подходов: zero-shot learning и few-shot learning. К сожалению, используя первый подход нам не удалось получить желаемых результатов, тем не менее во втором случае мы получили довольно хорошую выдачу, генерируемую моделями T5 и Flan-T5.

Для получения еще более высоких результатов мы также провели несколько оценочных экспериментов, а именно добавление дополнительных примеров на вход модели во few-shot подходе и дообучение моделей. Первая стратегия оказалась неудачной и результаты даже немного ухудшились, тем не менее второй подход позволил достичь того, что дообученные модели T5 и Flan-T5 практически перестали допускать ошибки в распознавании объектов и аспектов сравнения в вопросительных предложениях. Дополнительные оценки работы дообученных моделей, помимо оценок, которые представлены в разделе Методология, можно увидеть ниже:

Таблица 22. Оценка работы дообученных моделей T5 и Flan-T5: precision, recall, f-score

		Precision	Recall	F-score
T5	Objects	0.724	1.0	0.839
	Aspect	0.759	0.909	0.828
Flan-T5	Objects	0.738	1.0	0.849
	Aspect	0.771	0.901	0.831

6.2. Оценка генерации суммаризации

Вторая задача состояла в представлении результатов выдачи в более кратком и понятном для пользователя формате, а именно в виде суммаризации полученных с помощью системы САМ аргументов в пользу каждого из объектов. Для выполнения этой задачи мы также решили использовать генерацию с помощью генеративных

моделей, а именно GPT-2 и GPT-3. Вторая модель использовалась в качестве генератора набора данных с примерами ввода и выдачи, часть из которых в последствии была проверена на наличие логических ошибок в реферировании и обработана нами вручную. GPT-2 же мы использовали в двух версиях – более легкой на 124М параметрах и средней по тяжести, а именно 345М.

Каждая из версий была дообучена в ходе нескольких экспериментов, целью которых являлось выявить наиболее успешную стратегию обучения для качественной суммаризации полученных данных в удобном формате. Каждый эксперимент включал в себя автоматическую оценку с помощью метрик ROUGE и BLEU, которые оценивали схожесть сгенерированного текста моделью GPT-3 и дообученными версиями модели GPT-2.

В ходе первых двух экспериментов модели обучались на 50 и 40 отредактированных примерах соответственно, но в первом случае автоматическая оценка указывала на схожесть сгенерированных текстов с необработанными результатами, а во втором, напротив, с исправленными данными. Оценка сравнения с исправленными данными оказалась выше, хоть и незначительная, что может говорить о том, что модель все же генерирует более логичную и качественную суммаризацию, если обучается на более хороших данных.

Следующие два эксперимента были направлены на расширение обучающей выборки двумя способами: добавление к 50 размеченным примерам 50 тех же самых без разметки или же 50 совершенно отличающихся. Мотивация такого решения описана в разделе Методология. Оценки, за исключением средней длины генерируемого текста, практически не отличаются. Генерируемая же длина напрямую зависит от разброса длины текстов во всем наборе данных.

Последний же эксперимент отличался от всех предыдущих тем, что мы сделали упор на количестве данных, а не на их качестве, поскольку третий и четвертый эксперименты показали более хорошие результаты, чем первый и второй. В ходе данного эксперимента мы выяснили, что размер обучающей выборки влияет на качество генерации модели сильнее, чем качество, поскольку даже при оценке на размеченных примерах результаты оказались наилучшими.

Помимо качественно сгенерированной суммаризации мы планировали добавить в текст реферирование к примерам из выдачи САМ, тем не менее модель не справилась с использованием ссылок, поскольку при ручной оценке GPT-2 показала лишь 0.36 точности в соотношения высказываний.

6.3. Возможные улучшения

В начале своего исследования мы планировали создать демонстрационную версию системы в виде веб-приложения. К сожалению, на данный момент нам не удалось выполнить эту задачу, тем не менее мы хотели бы предложить вариант ее исполнения.

Инструмент предлагается разделить на две логические части: с пользовательским вводом и результатами выдачи. Вводные данные можно принимать в двух форматах, в классическом формате, используемом в демоверсии САМ, а именно с полями для эксплицитного ввода объектов и аспекта сравнения, или же в виде поля для ввода полного вопроса на естественном языке. Более того, необходимо добавить возможность выбирать количество высказываний о каждом объекте, на основе которых будет производиться суммаризация.

Мы бы также хотели предоставить пользователю выдачу в двух форматах: в виде суммаризации, сгенерированной моделью, а также в виде списка предложений, на основе которых она сделана. Для более общих результатов можно добавить оценочную шкалу, как это представлено в демоверсии САМ.

Макет планируемого веб-интерфейса можно подробнее изучить в Приложении 2.

7. Заключение

Данное исследование представляет собой создание системы ответов на сравнительные вопросы, содержащие в себе два объекта и аспект сравнения. Наш инструмент является расширенной версией системы Comparative Argumentative Machine (или САМ), которая производит поиск и оценку сравнительных предложений с искомыми объектами и аспектом сравнения на основе данных из масштабного веб-архива. Мы расширили работу САМ с помощью генеративных моделей, решающих задачи распознавания объектов и аспекта сравнения в вопросительном предложении на естественном языке, а также генерации ответа в виде суммаризации данных выдачи САМ.

В ходе исследования нам удалось успешно обучить три генеративных модели: T5 и Flan-T5 для первой задачи и GPT-2 в двух версиях для второй, первые из которых с высокой точностью выделяют нужную информацию из вопроса и практически не допускают ошибок в выделении объектов и аспекта сравнения, а третья генерирует

качественный логичный текст суммаризации данных при небольших затратах вычислительной мощности.

Все используемые нами в ходе исследования модели были тщательно оценены, также мы привели описание множества экспериментов и стратегий обучения генеративных моделей. Мы надеемся, что полученные нами результаты помогут будущим исследованиям, а представленная система будет иметь дальнейшее развитие в научной среде.

Приложения

График 1. Сравнение машинного и профессионального ручного переводов по данным eff.org

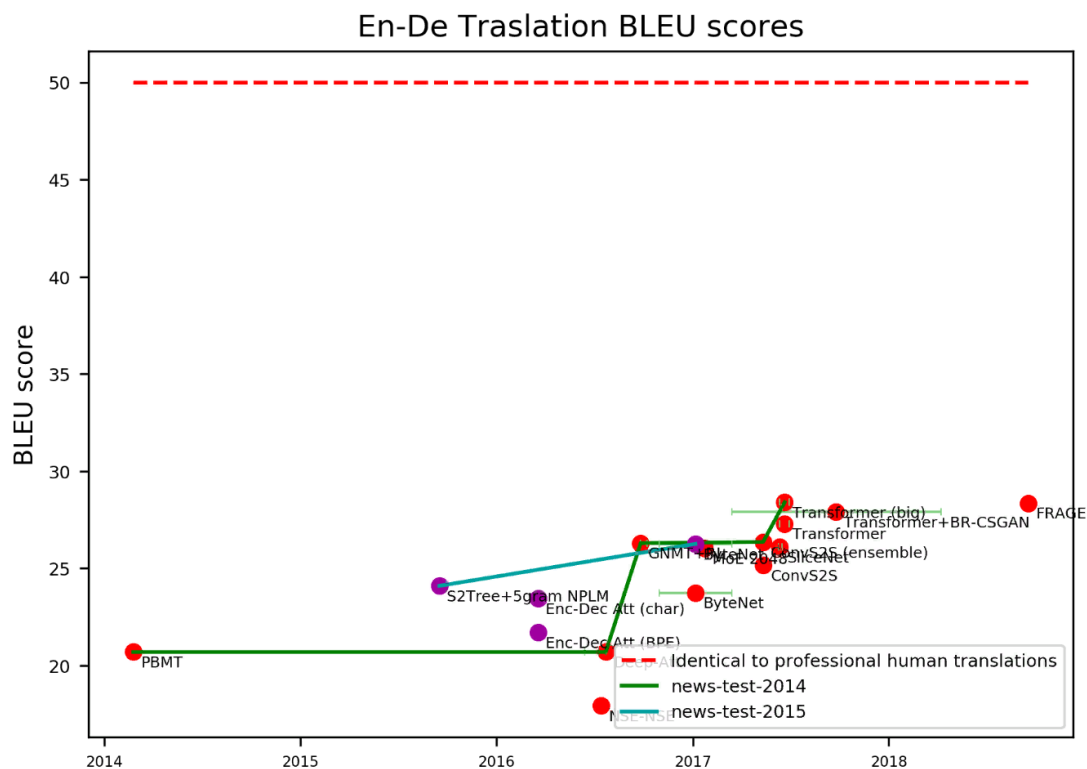


Рисунок 2. Макет веб-интерфейса

Comparative Argumentative Machine v2

Enter your question here...

or enter two objects for comparison

Object 1
milk

Object 2
soda

Aspect
taste

number of arguments for each object
10

Compare

milk (68%)

soda (32%)

Summary

Milk and soda are two popular beverages with different nutritional values and health implications.

Many argue that milk is better than soda in terms of nutritional value [1][8][12] and as a healthier alternative [3][5][6]. Some even argue that flavored milk is worse than soda, as it contains more sugars and corn syrups [9]. People also prefer the taste of milk over sugary soda and juice [3][4].

However, there are also arguments against milk, as it is said to be worse for health than soda [10][11]. Some people have lactose intolerance and cannot consume milk without digestive discomfort. Milk jugs are also harder to seal and degrade quicker than soda bottles [2][7].

Ultimately, the decision between milk and soda depends on personal preferences. Some might prefer milk for its nutritional benefits, while others might choose soda for its taste and convenience. As always, moderation is key, and it's a good idea to balance the consumption of both beverages with a healthy diet and exercise.

Sources

- [1. even homogenized pasteurized milk is far better then soda.](#)
- [2. Plastic milk jugs are hard to seal and degrade quicker than plastic soda bottles.](#)
- [3. Milk is way better than sugary soda, sweet chocolate milk, or orange juice.](#)
- [4. I suppose milk is better than Kool-Aid and soda.](#)
- [5. Milk is a much better alternative than a can of soda.](#)
- [6. Milk & water are just plain better than soda, tea or juice.](#)
- [7. I like using the milk jugs better than the soda bottles.](#)
- [8. I rank pasteurized milk slightly better than soda in nutritional value.](#)
- [9. Flavored milk is even worse, as it contains corn syrups and sugars that make is more like soda than milk.](#)
- [10. Milk, although it might offer some nutritional value, is overall worse for your health than soda.](#)
- [11. A long time ago, doctors endorsed cigarettes, and soda was better than mother's milk.](#)
- [12. Lactose, or milk sugar, is metabolized, or broken down and used by the body, at a slower rate than say, soda.](#)

GitHub с исходным кодом: <https://github.com/ra4wv2/CompQA>

Литература

Arora et al. 2017 – Arora J. et al. Extracting entities of interest from comparative product reviews //Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. – 2017. – C. 1975-1978.

Beloucif et al. 2022 – Beloucif M. et al. Elvis vs. M. Jackson: Who has More Albums? Classification and Identification of Elements in Comparative Questions //Proceedings of the Thirteenth Language Resources and Evaluation Conference. – 2022. – C. 3771-3779.

Black et al. 2021 – Black S. et al. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow //If you use this software, please cite it using these metadata. – 2021. – T. 58.

Bondarenko et al. 2020a – Bondarenko A. et al. Comparative web search questions //Proceedings of the 13th International Conference on Web Search and Data Mining. – 2020. – C. 52-60.

Chekalina et al. 2021 – Chekalina V. et al. Which is better for deep learning: python or MATLAB? Answering comparative questions in natural language //Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. – 2021. – C. 302-311.

Chung et al. 2022 – Chung H. W. et al. Scaling instruction-finetuned language models //arXiv preprint arXiv:2210.11416. – 2022.

Devlin et al. 2019 – Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.

Floridi, Chiriatti 2020 – Floridi L., Chiriatti M. GPT-3: Its nature, scope, limits, and consequences //Minds and Machines. – 2020. – T. 30. – C. 681-694.

Gao et al. 2020 – Gao L. et al. The pile: An 800gb dataset of diverse text for language modeling //arXiv preprint arXiv:2101.00027. – 2020.

Lin 2004 – Lin C. Y. Rouge: A package for automatic evaluation of summaries //Text summarization branches out. – 2004. – C. 74-81.

Liu et al. 2022 – Liu Y. et al. BRIO: Bringing order to abstractive summarization //arXiv preprint arXiv:2203.16804. – 2022.

Liu et al. 2019 – Liu Y. et al. Roberta: A robustly optimized bert pretraining approach //arXiv preprint arXiv:1907.11692. – 2019.

Luo 2005 – Luo X. On coreference resolution performance metrics //Proceedings of human language technology conference and conference on empirical methods in natural language processing. – 2005. – C. 25-32.

Mallick et al. 2019 – Mallick C. et al. Graph-based text summarization using modified TextRank //Soft Computing in Data Analytics: Proceedings of International Conference on SCDA 2018. – Springer Singapore, 2019. – C. 137-146.

Panchenko et al. 2017 – Panchenko A. et al. Building a web-scale dependency-parsed corpus from CommonCrawl //arXiv preprint arXiv:1710.01779. – 2017.

Papineni et al. 2002 – Papineni K. et al. Bleu: a method for automatic evaluation of machine translation //Proceedings of the 40th annual meeting of the Association for Computational Linguistics. – 2002. – C. 311-318.

Peters et al. 2018 – Peters M. E. et al. Deep contextualized word representations. CoRR abs/1802.05365 (2018) //arXiv preprint arXiv:1802.05365. – 1802.

Radford et al. 2019 – Radford A. et al. Language models are unsupervised multitask learners //OpenAI blog. – 2019. – T. 1. – №. 8. – C. 9.

Raffel et al. 2020 – Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer //The Journal of Machine Learning Research. – 2020. – T. 21. – №. 1. – C. 5485-5551.

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

Reimers, Gurevych 2020 – Reimers N., Gurevych I. Making monolingual sentence embeddings multilingual using knowledge distillation //arXiv preprint arXiv:2004.09813. – 2020.

Schildwächter et al. 2019 – Schildwächter M. et al. Answering comparative questions: Better than ten-blue-links? //Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. – 2019. – C. 361-365.

Sundheim, Sundheim 1996 – Grishman R., Sundheim B. M. Message understanding conference-6: A brief history //COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. – 1996.

Sutton, McCallum 2012 – Sutton C. et al. An introduction to conditional random fields //Foundations and Trends® in Machine Learning. – 2012. – T. 4. – №. 4. – C. 267-373.

Touvron et al. 2023 – Touvron H. et al. Llama: Open and efficient foundation language models //arXiv preprint arXiv:2302.13971. – 2023.

Vaswani et al. 2017 – Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – T. 30.

Wang, Komatsuzaki 2021 – Wang B., Komatsuzaki A. GPT-J-6B: A 6 billion parameter autoregressive language model. – 2021.