Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

КУРСОВАЯ РАБОТА

Исследовательский проект на тему

"Разработка бенчмарка RussianSuperGLUE"

Выполнил студент группы 172, 4 курса,
Кириллов Дмитрий Александрович

Руководитель КР:
канд. техн. наук, доцент, Артемова Екатерина Леонидовна

Москва 2021

Оглавление

Аннотация

В последние несколько лет предварительно обученные нейросетевые языковые модели находят все большее применение в различных задачах обработки естественного языка. Для оценки и сравнения таких моделей на русском языке активно разрабатывается бенчмарк RussianSuperGLUE. На данный момент он содержит меньшее число заданий, чем аналоги на английском языке. В рамках данной работы был разработан набор данных на русском языке, который позволяет оценивать качество решения задачи обнаружения парафраза.

Ключевые слова— Обнаружение парафраза, обработка естественного языка, GLUE, разработка датасета, QQP, перенос обучения

Abstract

In the past few years, pretrained neural language models have found broad application in various natural language processing tasks. RussianSuperGLUE is an actively developed benchmark for evaluating and comparing such Russian-language models. Nevertheless, he is currently inferior to counterpart in English in the number of tasks. The paper introduces a question pairs dataset in Russian that can be used to evaluate the quality of the paraphrase detection model.

1 Введение

1.1 Описание предметной области

Бенчмарк в контексте машинного обучения — набор заданий, позволяющих оценить качество модели. Каждое задание соответствует конкретной задаче машинного обучения и представлено выборкой данных, характерных для данной задачи. В задании обязательно присутствует недоступная во время обучения тестовая выборка, на которой и измеряется качество ответов модели с помощью заранее определенных метрик качества. Обычно задание также содержит обучающую выборку. Однако в процессе обучения многих моделей зачастую используются дополнительные данные.

Для задач обработки текстов на английском языке существует большое количество бенчмарков. GLUE [21] содержит набор заданий, которые позволяют оценить общее понимание естественного языка моделью. Данная оценка наиболее актуальна для современных моделей на основе архитектуры «Трансформер» [19], которые демонстрируют впечатляющие результаты во многих задачах обработки естественного языка.

Более того, качество решения предложенных задач новейшими языковыми моделями вплотную приближается или даже превосходит результат, демонстрируемый человеком. Бенчмарк SuperGLUE [20] предлагает набор более сложных и разнообразных заданий.

Аналогичный подход к оценке качества языковых моделей для русского языка реализует бенчмарк RussianSuperGLUE [18]. На данный момент в нем отсутствуют многие задания, представленные в англоязычных аналогах.

Одной из таких задач является обнаружение парафраза — пары различных предложений, совпадающих по смыслу. Для оценки качества определения парафраза на английском языке наибольшую популярность получил набор данных «Quora Question Pairs» («QQP») [8]. Он состоит из пар коротких вопросов. Каждой паре сопоставлена бинарная метка, показывающая повторяют ли вопросы друг-друга по смыслу.

1.2 Постановка задачи

В работе была поставлена задача создания набора данных, аналога «QQP» на русском языке, применимого для оценки качества обнаружения парафраза предварительно обученными языковыми моделями. Набор должен состоять из пар коротких вопросов. Каждой паре необходимо сопоставить бинарную метку, показывающую является ли она парафразом. Разметка должна быть произведена полностью вручную с помощью краудсорнинговой платформы.

2 Обзор литературы

Задача обнаружения парафраза между парой документов является распространенной задачей анализа естественного языка. Как следствие существуют несколько наборов данных, состоящих из пар семантически близких документов, различающихся по размеру, специфике текстов и подходу к сбору и разметке данных.

Наиболее объемные наборы получены с помощью автоматического составления парафразов. Авторы набора «PPDB» [7] представили корпус из 230 миллионов пар на английском языке, полученный с использованием моделей семантической близости предложений. Набор «ParaNMT» [22] из 50 миллионов пар был создан из параллельного корпуса текстов на английском и чешском языках с помощью предварительно обученной нейросетевой модели машинного перевода текстов. Хотя процесс автоматической разметки данных может приводить к появлению ошибок в полученных парах, авторы демонстрируют, что полученные наборы могут успешно применяться для обучения моделей обнаружения парафразов.

Однако подобные наборы данных не подходят для оценки качества языковых моделей, так как они могут содержать значительное число неточно размеченных пар. Как следствие значение метрик качества, вычисленное на основе подобных выборок, не будет отражать реальное качество модели. Поэтому для оценки и сравнения моделей выявления парафраза обычно используют вручную размеченные данные. Наиболее популярным является корпус «QQP» [8], составленный из пар вопросов, опубликованных пользователями сайта Quora.com, и размеченный вручную. Несмотря на то, что авторы не публикуют подробную методику составления корпуса и признают возможность наличия ошибок в нем, этот набор данных успешно применяется для оценки качества моделей.

При комбинировании техник ручной разметки и автоматической генерации итоговый корпус получается значительно большего размера в сравне-

нии с наборами, размеченными исключительно вручную, и содержит меньше ошибок, чем наборы данных, составленные полностью автоматически. Составители набора «Language-Net» [12] демонстрируют, что модель, обученная на вручную размеченном наборе данных, позволяет в дальнейшем непрерывно отбирать новые пары-кандидаты с точностью близкой к 70%.

В другой работе [23] был предложен подход к расширению существующего корпуса парафразов за счет генерирования новых пар, полученных путем перестановки слов в исходных предложениях. Авторами был сделан вывод о невысокой точности данного способа из-за чувствительности предложений на английском языке к изменению порядка слов. Однако данный подход может оказаться более эффективным при применении к корпусу на русском языке, в котором перестановка слов менее выражено изменяет смысл предложения.

Один из немногих корпусов парафраза на русском языке «Paraphraser» [15] является полностью размеченным вручную. Предложения в корпусе получены из заголовков публикаций российских новостных агентств. Для отбора пар-кандидатов авторы используют собственную метрику близости предложений, что позволяет повысить эффективность дальнейшей ручной обработки. В отличие от ранее рассмотренных, в представленном наборе данных пары делятся на три класса: полностью совпадающие по смыслу (1), близкие по смыслу (0) и разные по смыслу (-1). Авторы указывают, что из-за специфики процесса разметки набор содержит мало примеров пар отрицательного класса.

Для решения поставленной задачи был выбран подход, предполагающий включение в корпус только вручную классифицированных пар, предварительно отобранных с использованием модели семантической близости предложений. Так как корпус «Quora Question Pairs» успешно применяется для оценки качества моделей выявления парафраза на английском языке, было решено использовать аналогичный источник данных и структуру классов.

3 Метод сбора данных

3.1 Сбор корпуса вопросов

На русском языке существуют два крупнейших ресурса в сети Интернет, предоставляющих пользователям возможность публиковать свои и отвечать на чужие вопросы: «Yandex Q» и «Ответы Mail.ru». Оба ресурса содержат большое количество вопросов, которые соответствуют критериям, предъявляемых к источнику текстов в рамках задачи. Для выбора одного из них был проведен анализ пользовательских соглашений.

В пользовательском соглашении «Ответы Mail.ru» [25] был указан прямой запрет на использование средств для автоматизации взаимодействия с ресурсом.

Правила пользования сервисом «Yandex Q» [26] в период, когда проводился сбор корпуса, подобного ограничения не содержали. Более того, соглашение закрепляет за ООО «ЯНДЕКС» возможность предоставить право третьим лицам хранить и распространять размещенные пользователями материалы.

Поэтому в качестве источника для сбора данных был выбран сайт «Yandex Q». Так как соглашение об использовании сервиса сохраняет авторское право на публикуемые вопросы за разместившими их пользователями, то при решении поставленной задачи непосредственно тексты вопросов не сохранялись и не публиковались. При необходимости провести анализ, текст каждый раз временно загружался со страницы вопроса. На некоторых страницах портала явно указана одна или несколько тем, к которым относиться вопрос. Эти темы не являются уникальным пользовательским материалом, поэтому собирались и сохранялись нами как дополнительная информация.

С помощью библиотеки `Scrapy` [17] на языке Python рекурсивно обходились страницы сайта. Когда посещалась страница, содержащая вопрос, сохранялась ссылка на нее и темы, ассоциированные с вопросом. Данные сохранялись в SQL базу данных. Идентификатор полученной записи использовался как уникальный номер соответствующего вопроса на всех этапах дальнейшей

обработки.

Ссылки собирались непрерывно в течении 19 дней. Всего было собрано 267 640 ссылок на вопросы. 248 278 из них относились к определенной теме.

3.2 Отбор кандидатов на разметку

Ручная разметка всех возможных пар такого объемного корпуса не представлялась возможной в рамках данной работы. Поэтому для дальнейшей разметки были отобраны лишь некоторые пары вопросов.

В первую очередь в корпусе были оставлены только вопросы, тексты которых соответствовали следующим критериям:

- Содержит только символы ASCII или кириллические символы Unicode
- Не содержит обсценных слов
- В тексте не встречаются слишком редкие слова (имеющие статистику IPM < 2 по частотному словарю [24])
- Состоит не менее чем из 6 слов
- Состоит не более чем из 12 слов

На последней стадии отбора пар-кандидатов были получены векторные представления всех предложений с использованием модели SentenceBERT [16], предварительно обученной на корпусе русского языка лабораторией DeepPavlov [3]. Данная модель обучалась авторами таким образом, чтобы для схожих по смыслу предложений их векторные представления имели большое значение косинусной близости.

При ручной разметке небольшого числа случайно выбранных пар не было обнаружено парафразов, у которых векторные представления имели близость ниже 0.9. При этом в группе с большим значением близости доля парафразов была примерно равна 0.2. Можно заключить, что современные предварительно обученные языковые модели способны выявлять близкие по смыслу пред-

ложения без дополнительного обучения под конкретную задачу. Однако для таких моделей может быть затруднительно определить пары предложений, состоящих из похожих слов и близких синтаксически, но разных по смыслу.

Для дальнейшей разметки были выбраны пары, у которых косинусная близость между полученными векторами оказалась не меньше 0.9.

3.3 Разметка отобранных пар

Для получения метки класса было составлено задание на краудсорсинговой платформе «Яндекс Толока». Платформа предоставляет возможность получить решение набора специально оформленных, обычно простых, заданий. Задания доступны для выполнения на возмездной основе большому числу пользователей.

Исполнителям предлагалось проследовать по двум гиперссылкам, указывающим на страницы вопросов, и дать ответ являются ли два вопроса одинаковыми по смыслу. Каждая пара была размечена как минимум 3 разными исполнителями.

Для улучшения качества итоговой разметки были использованы несколько способов контроля, доступных на платформе. Наиболее важными из них были: обязательное прохождение обучения перед выполнением заданий и контрольные вопросы.

Обучение представляло собой набор характерных пар с известными метками класса. Если исполнитель давал неверный ответ на задание, ему демонстрировался подробный комментарий, аргументирующий выбор правильного ответа.

Контрольные вопросы — задания, для которых известен правильный ответ, но неотличимые для исполнителя от обычного задания. Было решено сделать 20% всех заданий контрольными. Для составления коллекции необходимого размера была применена следующая процедура:

1 Нами собственноручно было размечено 100 контрольных заданий.

2 Была получена разметка для 500 пар.

3 Отобраны 300 пар с наибольшей согласованностью ответов исполнителей, с помощью модели [2], доступной на платформе.

4 Эти пары вместе с исходными были использованы для составления контрольных заданий.

В рамках доступного бюджета удалось получить разметку для 1915 пар. Валидационная выборка опубликована [1] в открытом доступе.

Таблица 3.1: Пример итоговой разметки пар

left_id	left_url	right_id	right_url	class
60130	https://yandex.ru/q/question/business/kakie_dokumenty_nuzhny_dlia_prodleniia_60766740/	243915	https://yandex.ru/q/question/kakie_dokumenty_nuzhny_dlia_polucheniia_c208b2c3/	0
73372	https://yandex.ru/q/question/kakaia_samaia_smeshnaia_kniga_kotoruiu_vy_1d1f4408/	93744	https://yandex.ru/q/question/kakuiu_samuiu_smeshnuiu_knigu_vy_chitali_1bfd9c12/	1
28539	https://yandex.ru/q/question/transport/kak_chasto_nuzhno_meniat_vozdushnyi_filtr_f0c703a3/	66554	https://yandex.ru/q/question/transport/kak_chasto_nado_meniat_tormoznuiu_b7acdca5/	0

4 Анализ собранных данных

Количество размеченных пар получилось существенно меньше, чем у аналогичных наборов, рассмотренных ранее. Так как изначально была поставлена задача создания набора данных для оценки качества и сравнение предварительно обученных языковых моделей, то было решено исследовать возможность применения собранных данных только в виде тестовой и валидационной выборки.

Для получения итоговой метки классов из ответов нескольких исполнителей применялась модель [2], использующая EM-алгоритм для максимизации

[1] https://github.com/wosadeh/Russian-Question-Pairs/raw/9b23240454356d1aaadf81a55f5b2847dec0b415/dev.tsv

правдоподобия итоговой разметки. Данная модель помимо самой метки оценивает уверенность в агрегированном значении. Из набора были исключены размеченные пары вопросов, уверенность на которых была ниже 92%.

В тестовую выборку были включены по 300 примеров отрицательного и положительного класса с уверенностью в метке не менее 99%.
В валидационную выборку вошли все остальные пары вопросов.

4.1 Описание данных

Для обеих выборок была вычислена мера [6] межэкспертной надежности полученной разметки. Чтобы иметь представление о сложности текстов, были вычислены статистики: метрика MTLD [13] лексического разнообразия, среднее количество слов в вопросе, разница в длине между вопросами в одной паре. Также с помощью библиотеки `pymorphy2` [10] были выявлены некоторые морфологические свойства текстов. Данные представлены в таблице 4.1.

Так как при разметке использовались контрольные задания, то для них

Таблица 4.1: Описательные статистики

Статистика \ **Подвыборка**	Тестовая	Валидационная
Размер	600	1121
Доля примеров положительного класса	0.5	0.24
Межэкспертная надежность	0.77	0.63
MTLD	25.1	31.9
Среднее количество слов в вопросе	8.03	8.09
Средняя разница количества слов в паре вопросов	1.3	1.4
Доля пар, содержащих числа	0.115	0.145
Доля пар, содержащих деепричастие	0.031	0.032
Доля пар, содержащих отрицание	0.085	0.105

возможно посчитать среднюю долю верных ответов. Эту величину, равную 84.34%, можно рассматривать как приблизительную оценку качества определения парафраза людьми в примерах из нашего корпуса.

Полученные значения для межэкспертной надежности свидетельствуют о высокой согласованности ответов исполнителей для пар, вошедших в итоговые выборки.

В двух подвыборках значительно отличаются доли положительных пар. Преимущество сбалансированной тестовой выборки состоит в более показательной оценке при использовании некоторых метрик качества, чувствительных к балансу классов. Несбалансированная валидационная выборка позволяет использовать большее количество размеченных данных. Это различие также проявляется и в большем значении метрики MTLD.

Остальные рассмотренные статистики имеют близкое значение у обеих выборок.

4.2 Тестирование моделей

Чтобы оценить качество языковых моделей на полученной тестовой выборке были рассмотрены несколько техник переноса обучения с других наборов данных, схожих с полученным нами.

4.2.1 Paraphraser

В первом эксперименте была использована предварительно обученная для русского языка модель RuBERT [11]. Архитектура BERT, на которой она основана, изначально поддерживает обработку пары предложений, разделенных специальным токеном `[SEP]`. Контекстуальное векторное представление специального токена `[CLS]` передавалось на вход двухслойной полносвязной нейросети для классификации.

Для обучения использовался набор данных «Paraphraser» [15]. В этом корпусе пары предложений разделены на 3 класса. Для обучения модели бинарной классификации использовалась кросс-энтропийная функция потерь.

$$\mathcal{L}(p, y) = -q_y \log(p) - (1 - q_y) \log(1 - p) \tag{1}$$

Где p – возвращаемая моделью вероятность положительного класса, y – метка класса для пары предложений. q_y выбрана равным 1 для «точного парафраза», 0 для пар различных предложений и равным 0.5 для частично

совпадающих по смыслу пар.

В качестве оптимизатора использовался Adam [9] с параметрами $\beta_1 = 0.9$, $\beta_2 = 0.999$, скоростью обучения $5 \cdot 10^{-5}$, размером пакета равным 32.
В процессе обучения обновлялись веса только двух последних слоев BERT и полносвязные слои для классификации.
Обучение продолжалось 10 эпох. После каждой эпохи обучения с использованием валидационной выборки вычислялось значение метрики Accuracy. В качестве итоговой выбиралась промежуточная модель с наибольшим значением метрики.

4.2.2 Перевод QQP

Во втором эксперименте обучающие данные были получены с помощью машинного перевода языка корпуса «QQP» [8]. Для перевода использовалась предварительно обученная модель FairSeq [14].

Данная модель продемонстрировала [1] хорошее качество перевода с английского языка на русский как по метрике BLEU, так и по данной людьми оценке. Пример переведенных предложений приведен в таблице 4.2.

Таблица 4.2: Пример перевода

Исходное предложение	**Русский перевод**
What is the step by step guide to invest in share market in india?	Каково пошаговое руководство по инвестированию в рынок акций в Индии?
What are some examples of deuteromycota and how are they formed?	Каковы некоторые примеры теромикоты и как они формируются?
How light bend with gravity?	Как легко согнуть с гравитацией?

Используемая предобученная модель, архитектура и оптимизатор аналогичны прошлому эксперименту. В качестве функции потерь использовалась стандартная кросс-энтропия. Обучение длилось 3 эпохи.

4.2.3 LaBSE

Последний эксперимент проводился с другой предварительно обученной моделью LaBSE [5] имеющей сходную с BERT архитектуру. Данная мультиязычная модель обучалась авторами таким образом, чтобы предложения со

схожим смыслом имели близкие (в плане косинусного расстояния) векторные представления для `[CLS]`.

Для классификации пары предложений векторные представления токена `[CLS]` конкатенировались и подавались на вход полносвязной нейросети с двумя слоями.

Дополнительно, чтобы сохранить свойство исходной модели, для пар положительного класса в функцию ошибки было добавлено слагаемое, отвечающее косинусной близости векторных представлений. Функция потерь имела следующий вид:

$$\mathcal{L}(v_0, v_1, y) = -\log f_y(v_0, v_1) + \lambda \cdot y\,(1 - \phi(v_0, v_1)) \tag{2}$$

где y – метка класса, v_0, v_1 – векторные представления `[CLS]` токена для первого и второго предложения соответственно, $f_y(v_0, v_1)$ – выход последнего слоя нейросети для класса y, ϕ – косинусная близость двух векторов.

В эксперименте использовалась обучающая выборка «QQP» [8] на английском языке. Гиперпараметр λ выбран равным 0.1. Обучение продолжалось 3 эпохи с помощью оптимизатора Adam [9] с параметрами $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning_rate $= 10^{-5}$, коэффициентом L2 регуляризации равным 10^{-2} и пакетом размера 32.

4.2.4 Результаты

Каждый эксперимент повторялся 5 раз с различными инициализациями генератора псевдослучайных чисел. Как следствие в каждом запуске различались начальные веса последних слоев нейросетей и порядок обхода объектов выборки при стохастической оптимизации.

После обучения были вычислены метрики качества на тестовой выборке. Средние значения (μ) и стандартные отклонения (σ) приведены в таблице 4.3.

При использовании корпуса «Paraphraser» [15] в качестве обучающей выборки значения метрик Accuracy и F1 получились значительно ниже, чем

Таблица 4.3: Результаты тестирования

Эксперимент	Accuracy		F1		ROC-AUC	
	μ	σ	μ	σ	μ	σ
Paraphraser + RuBERT	74.17	01.28	72.93	02.81	82.90	01.72
FairSeq EN-RU + QQP + RuBERT	79.10	02.12	78.98	02.23	86.98	01.24
— " — (сбалансированная вал. выборка)	79.70	01.05	79.35	01.70	87.20	00.66
QQP + LaBSE	75.77	01.79	76.76	02.19	83.81	01.45
RuBERT + Paraphraser (тест) [11]	84.99	00.35	87.73	00.26	—	—
BERT + QQP (тест) [4]	89.30	—	71.20	—	—	—

удалось достичь авторам той же модели на тестовой выборке из этого корпуса. Это можно объяснить тем, что этот корпус значительно отличается от собранного нами как по источнику текстов, так и по структуре классов.

При использовании для обучения набора данных «QQP» [8] значение всех метрик качеств получилось значительно лучше. Достигнутая в наших экспериментах величина F1 не сильно отличается по значению от полученной [4] авторами модели BERT для тестовой выборки «QQP». Все это может свидетельствовать о том, что полученный нами набор данных действительно имеет сходство с аналогом на английском языке.

Второй эксперимент был также повторен с использованием подмножества валидационной выборки с равной долей примеров обоих классов. Значение метрик качества оказалось выше, а их разброс меньше. Разница оказалась не слишком велика, так как валидационная выборка использовалась исключительно для отбора лучшей из промежуточных моделей во время обучения. Тем не менее важно учитывать, что использование сбалансированного подмножества валидационной выборки может улучшить значение метрик качества на тестовой выборке.

Значения имеют больший разброс, чем полученные в схожих экспериментах авторами русскоязычной модели [11]. Скорее всего причиной послужил небольшой по сравнению с аналогами объем полученного корпуса и использование техник переноса обучения.

Значение метрик Accuracy и F1 в наших экспериментах имеют близкие значения, так как в тестовой выборке доли пар обоих классов равны.

Из данных экспериментов видно, что в каждом из них примерно каждая пятая пара классифицируется моделью неверно. Возможно причина состоит в том, что на этапе отбора кандидатов на разметку выбирались те пары предложений, которые определялись SentenceBERT моделью как семантически близкие, при том что большая часть из них в итоге не была размечена людьми как парафраз. Такие примеры являются априори сложными для правильной классификации моделями из семейства BERT. Из описательных статистик для выборок видно, что примерно каждая десятая пара содержит числа или отрицание. Автоматический анализ подобных предложений также может быть затруднительным.

Можно сделать вывод, что полученный корпус возможно использовать для оценки качества обнаружения парафраза предварительно обученными моделями. Однако полученная оценка будет значительно зависеть от выбранной техники переноса обучения. Поэтому для корректного сравнения различных моделей необходимо зафиксировать такую технику или дополнить полученный корпус обучающей выборкой в рамках дальнейших исследований.

Исходный код программ, использованных для сбора, обработки данных и проведения экспериментов, доступен в репозитории[2].

5 Заключение

В рамках работы был создан корпус коротких вопросов-парафразов на русском языке. Полученный набор данных отличается от существующих аналогов для русского языка.

Отобранные выборки данных возможно использовать для оценки качества обнаружения парафраза современными языковыми моделями при применении техник переноса обучения. Однако полученные при таком подходе значения метрик качества будут сильно зависеть от использованной техники.

Поэтому в рамках дальнейших исследований целесообразно разметить оставшиеся пары-кандидаты и выделить из них полноценную обучающую

[2] https://github.com/wosadeh/Russian-Question-Pairs

выборку.

Также возможно значительно улучшить формат представления данных, если получить согласие представителей сервиса «Yandex Q» на публикацию текстов вопросов.

Список литературы (или источников)

1. *Barrault L.*, *Bojar O.*, *Costa-Jussa M. R.*, *Federmann C.*, *Fishel M.*, *Graham Y.*, *Haddow B.*, *Huck M.*, *Koehn P.*, *Malmasi S.* [и др.]. Findings of the 2019 conference on machine translation (wmt19) // Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). — 2019. — с. 1—61.
2. *Dawid A. P.*, *Skene A. M.* Maximum likelihood estimation of observer error-rates using the EM algorithm // Journal of the Royal Statistical Society: Series C (Applied Statistics). — 1979. — т. 28, № 1. — с. 20—28.
3. DeepPavlov/rubert-base-cased-sentence. — 12.2020. — URL: https://huggingface.co/DeepPavlov/rubert-base-cased-sentence (дата обр. 01.02.2021).
4. *Devlin J.*, *Chang M.-W.*, *Lee K.*, *Toutanova K.* Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. — 2018.
5. *Feng F.*, *Yang Y.*, *Cer D.*, *Arivazhagan N.*, *Wang W.* Language-agnostic bert sentence embedding // arXiv preprint arXiv:2007.01852. — 2020.
6. *Fleiss J. L.* Measuring nominal scale agreement among many raters. // Psychological bulletin. — 1971. — т. 76, № 5. — с. 378.
7. *Ganitkevitch J.*, *Van Durme B.*, *Callison-Burch C.* PPDB: The paraphrase database // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2013. — с. 758—764.

8. *Iyer S.*, *Csernai K.*, *Dandekar N.* First Quora Dataset Release: Question Pairs - Data @ Quora. — 01.2017. — URL: https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs (дата обр. 01.02.2021).

9. *Kingma D. P.*, *Ba J.* Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. — 2014.

10. *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. т. 542 / под ред. M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets. — Springer International Publishing, 2015. — с. 320—332. — (Communications in Computer and Information Science). — URL: http://dx.doi.org/10.1007/978-3-319-26123-2_31.

11. *Kuratov Y.*, *Arkhipov M.* Adaptation of deep bidirectional multilingual transformers for russian language // arXiv preprint arXiv:1905.07213. — 2019.

12. *Lan W.*, *Qiu S.*, *He H.*, *Xu W.* A Continuously Growing Dataset of Sentential Paraphrases // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. — Copenhagen, Denmark : Association for Computational Linguistics, 09.2017. — с. 1224—1234. — URL: https://www.aclweb.org/anthology/D17-1126.

13. *McCarthy P. M.*, *Jarvis S.* MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment // Behavior research methods. — 2010. — т. 42, № 2. — с. 381—392.

14. *Ng N.*, *Yee K.*, *Baevski A.*, *Ott M.*, *Auli M.*, *Edunov S.* Facebook FAIR's WMT19 News Translation Task Submission // arXiv preprint arXiv:1907.06616. — 2019.

15. *Pivovarova L.*, *Pronoza E.*, *Yagunova E.*, *Pronoza A.* ParaPhraser: Russian paraphrase corpus and shared task // Conference on Artificial Intelligence and Natural Language. — Springer. 2017. — c. 211—225.

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

17. Scrapy 2.4 documentation. — 10.2020. — URL: https://docs.scrapy.org/en/2.4/ (дата обр. 01.02.2021).

18. *Shavrina T.*, *Fenogenova A.*, *Emelyanov A.*, *Shevelev D.*, *Artemova E.*, *Malykh V.*, *Mikhailov V.*, *Tikhonova M.*, *Chertok A.*, *Evlampiev A.* RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // arXiv preprint arXiv:2010.15925. — 2020.

19. *Vaswani A.*, *Shazeer N.*, *Parmar N.*, *Uszkoreit J.*, *Jones L.*, *Gomez A. N.*, *Kaiser L.*, *Polosukhin I.* Attention is all you need // arXiv preprint arXiv:1706.03762. — 2017.

20. *Wang A.*, *Pruksachatkun Y.*, *Nangia N.*, *Singh A.*, *Michael J.*, *Hill F.*, *Levy O.*, *Bowman S. R.* Superglue: A stickier benchmark for general-purpose language understanding systems // arXiv preprint arXiv:1905.00537. — 2019.

21. *Wang A.*, *Singh A.*, *Michael J.*, *Hill F.*, *Levy O.*, *Bowman S. R.* GLUE: A multi-task benchmark and analysis platform for natural language understanding // arXiv preprint arXiv:1804.07461. — 2018.

22. *Wieting J.*, *Gimpel K.* ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations // arXiv preprint arXiv:1711.05732. — 2017.

23. *Zhang Y.*, *Baldridge J.*, *He L.* PAWS: Paraphrase adversaries from word scrambling // arXiv preprint arXiv:1904.01130. — 2019.

24. *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка: на материалах Национального корпуса русского языка. — Азбуковник, 2009.

25. Пользовательское соглашение "Ответы Mail.ru". — 11.2018. — URL: https://help.mail.ru/legal/terms/answers/ua (дата обр. 26.12.2020).

26. Условия использования сервиса Яндекс.Кью. — 12.2019. — URL: https://yandex.ru/legal/q_termsofuse/ (дата обр. 26.12.2020).