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

Содержание

1 Аннотация

2 Введение

2.1 Описание предметной области

Бенчмарк в контексте машинного обучения – набор заданий, позволяющих оценить качество модели. Каждое задание соответствует конкретной задачи машинного обучения и представлено выборкой данных, характерных для данной задачи. Для каждого задания определена метрика качества.

Для задач обработки текстов на английском языке существует большое количество бенчмарков. GLUE [Wang, Singh (и др.), 2018] содержит набор заданий, которые позволяют оценить общее понимание языка моделью. Данная оценка наиболее актуальна для современных моделей на основе архитектуры «Трансформер», которые демонстрируют впечатляющие результаты во многих задачах обработки естественного языка.

Однако, качество решения предложенных задач последними языковыми моделями вплотную приближается или даже превышает результат, демонстрируемый человеком. Бенчмарк SuperGLUE [Wang, Pruksachatkun (и др.), 2019] предлагает набор более сложных и разнообразных заданий.

Аналогичный подход к оценки качества языковых моделей для русского языка реализует бенчмарк Russian SuperGLUE [Shavrina (и др.), 2020]. На данный момент данный бенчмарк содержит заметно меньшее число заданий, чем два вышеописанных. Поэтому существует необходимость добавления новых заданий.

2.2 Постановка задачи

В работе была поставлена задача создания набора данных с парами вопросов-парафразов на русском языке по аналогии с заданием на английском языке «Quora Question Pairs» [Iyer, Csernai, Dandekar, 2017].

Набор данных должен состоять из пар предложений на русском языке. Каждой паре необходимо сопоставить категориальную метку, показывающую

имеют ли предложения одинаковый смысл. Предложения должны представлять из себя вопросы, размещенные пользователями в сети Интернет на русскоязычном аналоге платформы Quora.com . Сопоставление каждой паре соответствующей метки должно производится с использованием краудсорнинговой платформы, предлагающей услуги ручной разметки данных.

Полученное множество необходимо разделить на обучающую и тестовую выборку.

3 Обзор литературы

Задача обнаружение парафраза между парой документов является распространенной задачей анализа естественного языка. Как следствие существуют несколько наборов данных, содержащих пары семантически близких документов, различающихся по размеру, специфики текстов и подходу к сбору и разметки данных.

Наиболее объемные наборы получены с помощью автоматического составления пар парафразов. В [Ganitkevitch, Van Durme, Callison-Burch, 2013] представлен корпус из 230 миллионов пар на английском языке, полученный с использованием моделей семантической близости предложений. Набор [Wieting, Gimpel, 2017] 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 использования при оценки качества языковых моделей, так как они содержат некорректные пары. Как следствие полученное значение метрики не будет отражать реальное качество модели. Поэтому для оценки и сравнения моделей выявления парафраза обычно используют вручную размеченные данные. Наиболее попу-

лярным является корпус [Iyer, Csernai, Dandekar, 2017], 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 данных, позволяет надежно оценить качество моделей.

При совместном использовании ручной разметки и моделей машинного обучения итоговый корпус получается значительно большего размера в сравнении с исключительно ручной разметкой и содержит меньше ошибок, чем набор данных, составленный автоматически. Составители используемого для сравнения моделей набора «Language-Net» [Lan (и др.), 2017] демонстрируют, что модель обученная на вручную размеченном наборе данных позволяет в дальнейшем непрерывно отбирать новые пары-кандидаты с точностью близкой к 70%.

В работе [Zhang, Baldridge, He, 2019] 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, так как перестановка слов менее выражено изменяет смысл предложения на русском языке.

Один из немногих корпусов парафраза на русском языке [Pivovarova (и др.), 2017] является полностью размеченным вручную. Предложения в корпусе получены из заголовков российских новостных агентств. Для отбора пар-кандидатов авторы используют собственную метрику близости предложений, что позволяет повысить эффективность дальнейшей ручной обработки. В отличии 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.

4 Метод сбора данных

4.1 Выбор источника

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 предложений для решаемой задачи. Для выбора одного из них был проведен анализ пользовательского соглашения и веб интерфейса.

4.2 Сбор данных

4.3 Предварительная фильтрация

4.4 Отбор кандидатов на ручную классификацию

4.5 Ручная разметка данных

5 Анализ собранных данных

Список литературы

Wang A., *Singh A.* [и др.]. GLUE: A multi-task benchmark and analysis platform for natural language understanding // arXiv preprint arXiv:1804.07461. — 2018.

Wang A., *Pruksachatkun Y.* [и др.]. Superglue: A stickier benchmark for general-purpose language understanding systems // arXiv preprint arXiv:1905.00537. — 2019.

Shavrina T. [и др.]. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // arXiv preprint arXiv:2010.15925. — 2020.

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).

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.

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

Lan W. [и др.]. 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.

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

Pivovarova L. [и др.]. ParaPhraser: Russian paraphrase corpus and shared task // Conference on Artificial Intelligence and Natural Language. — Springer. 2017. — с. 211—225.