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

Аннотация

В последние несколько лет предварительно обученные нейросетевые языковые модели находят все большее применение в различных задачах обработки естественного языка. Для оценки таких моделей на русском языке активно разрабатывается бенчмарк RussianSuperGLUE. В рамках данной работы решается задача создания набора коротких вопросов-парафразов на русском языке для расширения набора заданий бенчмарка.

Ключевые слова— Детекция парафраза, обработка естественного языка, GLUE, разработка датасета, QQP

Abstract

Pretrained neural language models find broad application in various tasks of natural language processing. RussianSuperGLUE benchmark offers a metric to measure and compare a quality of such models. In this paper we present a question paraphrase pairs dataset in Russian that can be used as new benchmark task.

Keywords— Paraphrase detection, NLP, GLUE, dataset development, QQP

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. Для каждого задания определена метрика качества.

Для задач обработки текстов на английском языке существует большое количество бенчмарков. GLUE [1] содержит набор заданий, которые позволяют оценить общее понимание естественного 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 новейшими языковыми моделями вплотную приближается или даже превышает результат, демонстрируемый человеком. Бенчмарк SuperGLUE [2] предлагает набор более сложных и разнообразных заданий.

Аналогичный подход к оценке качества языковых моделей для русского языка реализует бенчмарк RussianSuperGLUE [3]. На данный момент бенчмарк содержит заметно меньшее число заданий, чем два вышеописанных. Поэтому существует необходимость добавления новых заданий.

1.2 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» [4].

Набор данных должен состоять из пар предложений. Каждой паре необходимо сопоставить категориальную метку, показывающую имеют ли предложения одинаковый смысл. Предложения должны представлять собой вопро-

сы, 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. Для оценки полученных результатов необходимо исследовать возможность использование полученного корпуса для обучения современных моделей, сравнить объем и содержание корпуса с существующими аналогами.

2 Обзор литературы

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

Наиболее объемные наборы получены с помощью автоматического составления парафразов. В [5] представлен корпус из 230 миллионов пар на английском языке, полученный с использованием моделей семантической близости предложений. Набор [6] 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 содержат некорректные пары. Как следствие полученное значение точности не будет отражать реальное качество модели. Поэтому для оценки и сравнения моделей выявления парафраза обычно используют вручную размеченные данные. Наиболее популярным является

корпус [4], 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 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 набора «Language-Net» [7] демонстрируют, что модель, обученная на вручную размеченном наборе данных, позволяет в дальнейшем непрерывно отбирать новые пары-кандидаты с точностью близкой к 70%.

В работе [8] 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 [9] 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, предвари-

тельно 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

В качестве источника для сбора данных был выбран сайт «Yandex Q» в сети Интернет, на котором публикуются вопросы пользователей на русском языке. Так как правила пользования сервисом сохраняют 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 временно загружался со страницы вопроса.

С помощью библиотеки Scrapy 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. Для дальнейшего использования были отобраны вопросы, текст которых соответствовали следующим критериям:

- Содержит только символы ASCII или кириллические символы Unicode
- Не содержит обсценных слов
- В тексте не встречается слишком редких слов (чья статистика IPM ≥ 2 по частотному словарю [10])

- Состоит не менее чем из 6 слов
- Состоит не более чем из 12 слов

Для отбора пар-кандидатов были получены векторные представления всех предложений с использованием модели [11], предварительно обученной на корпусе русского языка лабораторией DeepPavlov. Для дальнейшей разметки были выбраны пары, у которых косинусная близость между полученными векторами оказалась больше заранее выбранного порога. Для оценки качества отбора кандидатов планируется построить диаграмму калибровки модели на основе небольшого количества размеченных пар с разной величиной косинусной близости.

Для получения метки класса было составлено задание на краудсорсинговой платформе «Яндекс Толока». В качестве элементов контроля качества работы пользователей, выполняющих разметку, каждая пара будет размечена 3 разными пользователями, оценены ответы пользователей на контрольных заданиях, введены ограничения для пользователей, отвечающих слишком быстро.

4 Анализ собранных данных

После агрегации ответов с «Яндекс Толоки», будет проведен анализ содержания и размера полученного корпуса и составлено его описание.

Для изучения возможности практического применения полученного корпуса будет дообучена модель [12] на обучающей подвыборке и измерена точность на тестовой выборке. Ожидается, что качество обученной модели будет сравнимо с качеством, достигаемым при использовании существующих корпусов парафраза.

Список литературы

1. *Wang A.*, *Singh A.* [и др.]. GLUE: A multi-task benchmark and analysis platform for natural language understanding // arXiv preprint arXiv:1804.07461. — 2018.
2. *Wang A.*, *Pruksachatkun Y.* [и др.]. Superglue: A stickier benchmark for general-purpose language understanding systems // arXiv preprint arXiv:1905.00537 2019.
3. *Shavrina T.* [и др.]. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // arXiv preprint arXiv:2010.15925. — 2020.
4. *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).
5. *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.
6. *Wieting J.*, *Gimpel K.* ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations // arXiv preprint arXiv:1711.05732. — 2017.
7. *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.
8. *Zhang Y.*, *Baldridge J.*, *He L.* PAWS: Paraphrase adversaries from word scrambling // arXiv preprint arXiv:1904.01130. — 2019.

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

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

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

12. *Devlin J.* [и др.]. Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. — 2018.