ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему Обратная языковая модель

Тема на английском Reversed Language Model

Студентка 2 курса

группы № 202

Зверева Арина Михайловна

Научный руководитель

Сериков Олег Алексеевич

Приглашенный

преподаватель

Москва, 2022 г.

Оглавление

1. Введение

В течение последних лет машинное обучение расширяется в особо быстром темпе и получает множество ветвей развития. Довольно важное место среди отраслей машинного обучения занимает обработка естественного языка (Natural Language Processing). Сейчас механизмы и алгоритмы, использующие данную обработку, можно встретить на каждом шагу, распознавание и обработка языка внедрились в повседневную жизнь (Siri, Cortana, Google Assistant и т. д.) Тем не менее, перед обработкой естественного языка появляется все больше задач, для решения которых создаются языковые модели.

В рамках данной курсовой работы будет разработана и описана обратная языковая модель. Задачей данной модели является предсказывание и написание начала текста по его продолжению. В отличии от типичных языковых моделей, которые интерпретируют контекстную информацию, данная модель должна понимать, что могло бы быть причиной данного контекста. Таким образом, языковая модель ранжирует не вероятности событий, а их причины. Это может быть полезно для решения задач, которые опираются на анализ причинно-следственных связей, и использовано для систем, направленных на обеспечение качества (Quality Assurance systems).

Мы начнем с обзора существующих подходов и моделей, которые в них используются. Затем перейдем к описанию использованных данных, а также отметим формат данных, ожидающийся нами на выходе. Далее будут описаны этапы процесса обучения модели, а также выбранные методы и алгоритмы. Мы рассмотрим поведение модели на нескольких задачах, а также проведем сравнение с базовой языковой моделью T5. Затем будет представлена лингвистическая интерпретация поведения моделей в сравнении, ее оценка и обсуждение результатов.

2. Обзор существующих моделей

Обучение моделей для обработки естественного языка получило значительное развитие и совершенство за последние годы. Возможно, это связано с использованием заранее подготовленных и обученных моделей, таких как BERT и T5 (Text-To-Text). Используя подобные модели, пользователь может сфокусироваться на корректировке и настройке модели на наборах меньших размеров для получения лучшего результата и большей эффективности.

Далее мы представим обзор языковой модели T5, так как в обучении нашей модели использовалась ее заранее обученная версия. Языковая модель T5 позволяет преобразовать все задачи по работе над обработкой естественного языка в единый формат, при котором исходные данные и данные на выходе являются текстовыми строками, что значительно упрощает работу с обучением модели.

Данная нейросетевая модель уже обучена многим задачам, затрагивающих работу с текстом. Например, перефразирование, заполнение пропусков, упрощение текста, диалоговый ответ, ответы на вопросы по содержанию текста и др. (Raffel et al. 2019)

Тем не менее, моделей, которые, наоборот, предсказывали бы начало текста по его продолжению, нам найти не удалось. Большинство моделей работают с текстом на основе ранжирования вероятности событий. В связи с этим, нам кажутся перспективными модели, опирающиеся на исследование причинно-следственных связей. Такие модели могут помочь решать задачи, направленные на понимание естественного языка (Natural Language Understanding). NLU является подразделом NLP, и проблемы, связанные с пониманием компьютером языка, еще только предстоит решать.

3. Данные (исходные данные и их формат, данные на выходе)

Для качественного и успешного обучения модели на каждом этапе требуется определенный набор данных. Условно данные можно разделить на две категории: обучающая выборка (training sample) и тестовая (контрольная) выборка (test sample).

Основной и самый продолжительный по времени этап обучения модели проводится на обучающей выборке. С помощью такой выборки производится настройка и отладка параметров и алгоритма. Модель учится на примерах, состоящих из исходных данных и ожидаемого результата на их основе.

Цель тестовой выборки заключается в оценке качества модели. По результатам такой выборки, можно понять, насколько хорошо модель натренировалась и стоит ли продолжать обучение, предлагая модели больший объем данных. Тестовая выборка не должна зависеть от обучающей для получения правдивой картины результатов.

Мы обучали модель на наборе данных Википедии, содержащем очищенны статьи на английском языке. Такой набор данных был использован моделью как тестовая выборка. Данный набор построен на основе дампа Википедии 2020 года и включает содержимое статей Википедии с очисткой, без ссылок, разметок, пустых разделов и т. д. Данные были взяты с сайта tensorflow, который предоставляет доступ к множеству датасетов. Наш

набор данных представляеввт собой набор из десяти файлов формата JSON, каждый из которых весит около двух гигабайтов.

Затем на основе этого набора данных был сделан датасет для дальнейшего обучения модели. Полученный нами датасет представляет собой словарь, в котором содержание статей разделено на данные, которые подаются при входе (input), и ожидаемый результат (target). В качестве исходных данных мы используем продолжение текста, а в качестве ожидаемого результата – то, что предшествовало. Таким образом, мы показываем модели то, чему она должна научиться.

4. Обучение модели

Код программы для получения и обработки данных и для обучения модели был написан на языке Python 3.7. Модель была инициализирована многозадачной версией с помощью библиотеки transformers. Обучение модели проходило на датасете, описанном выше. Далее модель совершенствуется на большем количестве данных с целью получения лучших результатов.

5. Поведение модели на нескольких задачах

В качестве контрольной выборки был использован benchmark superGLUE. К сожалению, модель не показала ожидаемых результатов и часто «мазала». Мы предполагаем, что это может быть связано с тем, что она обучалась на достаточно формальных статьях. SuperGLUE NLU benchmark отличается тем, что был специально разработан сложным для машинного обучения (Wang et al. 2019), и в связи с этим, кажется, что для результатов, приближенных к пониманию моделью естественного языка, необходимо дополнительное обучение.

6. Описание модели в сравнении с T5

Разработанная в рамках данной курсовой работы модель представляется трудной для сравнения с базовой T5. Наша модель уступает в скорости работы и производительности. Тем не менее, обратная языковая модель фокусируется на нахождении и изучении причинно-следственных связей, это отличает ее от базовой T5, фокус которой больше направлен на генерацию текста.

7. Лингвистическая интерпретация поведения моделей в сравнении

Нам кажется необходимым представить лингвистическую интерпретацию поведения моделей в сравнении. Во-первых, хотелось бы отметить, что в отличии от других моделей обратная языковая модель снимает фокус с задачи написания текста и переносит его на задачу понимания. В связи с этим такая модель имеет меньше творческой свободы для генерации текста, так как подобный порядок написания текста является неестественным для текстов, пишущихся человеком. Столкновение таких процессов приводит к ограничению друг друга в доступных вариациях. Нам кажется, что на основе этого такая модель сможет преуспеть в решении задач, связанных с пониманием естественного языка.

8. Заключение

Таким образом, можно сделать вывод, что задачи, связанные с обработкой естественного языка, благодаря использованию заранее обученных моделей, реализуются намного эффективнее. Уже сейчас модель T5 показывает невероятно успешные результаты на бенчмарках различной сложности.

Задачи в области понимания естественного языка еще предстоит решать, и мы надеемся, что модель, обученная в рамках данной курсовой работы, будет полезна для достижения таких задач. В качестве возможных перспектив мы хотели бы предложить дальнейшее обучение данной модели на датасетах, представляющих тексты разных стилей. Нам кажется, что исследование корреляции поведения модели и ее обучение понимания причинно-следственных связей представляет собой перспективную работу. Также интересным кажется, объединение обратной языковой модели и модели, направленной на предсказывание текста. Такая работа определенно затронет значимые моменты обучения языковых моделей.

9. Литература

Raffel et al. 2019 — C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Lei, P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // Journal of Machine Learning Research 21, 2019

Wang et al. 2019 – A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language

Understanding Systems // 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada

10. Приложение

https://www.tensorflow.org/datasets/catalog/overview - датасеты TensorFlow
https://github.com/google-research/text-to-text-transfer-transformer - T5
https://github.com/arinazv/reversed_lm – код программы и другие используемые файлы лежат по этому адресу