УДК 003.26

ОПРЕДЕЛЕНИЕ МЕСТА ДЕЙСТВИЯ НА ОСНОВЕ СЕМАНТИКИ ТЕКСТА.

Мицковский Д.Ю. Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

Научный руководитель – канд. физ.-мат. наук Фильченков А.А.

Научный консультант – Ефимова В.А.

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

В докладе представлены основные результаты исследования и реализации методики определения локаций текста на основе его семантики. Реализация основана на нейронных сетей с использованием трансформеров на основе BERT.

Введение.

Задача определения локации текста решается уже довольно давно и заключается она в поиске семантических связей между словами в тексте. В большинстве случае нужно понять, к какому слову относится то или иное существительное, глагол, прилагательное. Либо же, как в более узком случае, необходимо понять в каком месте происходит то или иное действие текста, то есть определить локацию действия на основе семантики. Решению задачи можно найти применение в создании машинных переводов текста, лингвистических стегосистемах, а также в переводе текстовой информации в графическую.

Поставленная задача считается актуальной, поскольку до сих пор не было получено универсального и надежного решения. Так, например, BERT[1] приходится «дообучать» при использовании в новом лингвистическом пространстве, а GPT-2[2] показывает плохие результаты на наборе данных COCO[3]. Актуальность исследований подтверждается наличием актуальных работ (2021 года), в которых демонстрируется сильный, но все же недостаточный прогресс в качестве выполнения.

Ранее для решения задачи была создана система из сиамских нейронных сетей BERT как объединение двух вариантов реализации сетей, а именно, на основе выделения локации в тексте (LET) и на основе вывода локации с помощью эмбеддинга, когда локация не дана сети (LIT). Однако они также не показали хороших результатов, но при этом показали значительный прогресс относительно других решений. Поэтому в рамках исследования необходимо выявить основные факторы, не позволяющие решениям достичь нужных результатов и сформировать теоретическое решение на основе выявленных факторов с учетом сильных и слабых сторон существующих решений.

Основная часть.

В рамках задачи обнаружения места действия текста на основе смысловых связей внутри самого текста объединенная сеть LET и LIT показала улучшение результатов на наборах данных. Однако неоднородность результатов на разных наборах данных, а также низкая точность LET на наборах данных вкупе с долгим временем выполнения и ограничением на входные данные. Поэтому основными направлениями работы являлись улучшение точности системы, уменьшения времени работы, а также сделать возможным работу с большими текстами и текстами на русском языке.

По итогам теоретического изучения проблем решения задачи было решено отойти от решения в виде сиамских сетей (в виду её долгой работы и недостаточной точности результатов) и была выбрана реализация на основе использования предобученной модели с помощью генератора на основе BERT (CodeBERT[4]). Генератор должен генерировать локации из наборов для входных данных (входных предложений). Сам генератор представляет из

[1] https://arxiv.org/pdf/1810.04805.pdf?source=post_elevate_sequence_page

[2] https://jalammar.github.io/illustrated-gpt2/

[3] https://cocodataset.org/#panoptic-2020

[4] https://arxiv.org/pdf/2002.08155.pdf

себя BERT, на вход которого и будут подаваться данные из набора. Дискриминатор, при этом, будет определять, является ли генерированная локация правдивой или нет.

Далее предобученная модель дообучается на наборе данных (данные поделены на две части: обучающую и валидирующую) с помощью сети с поитерационной валидацией точности на основе проверки работы сети. При каждой итерации сеть получает потери на обучении, а затем полученная модель пытается проставить метки (локации) на валидирующей части набора данных. Полученные потери на валидации, в дальнейшем, используются для дообучения сети.

При обучении модели на вход подаются предложения и локации, которые переводятся в токены с помощью токенизаторов. Далее токены идут на вход предобученной модели, полученной на генераторе. Затем, для подсчёта потерь при обучении используется CrossEntropyLoss на основе градиентов. После этапа обучения создаётся bin файл модели, а после, на основе дообученной модели получаются предсказания для валидрующей части набора данных. На основе полученных потерь на предсказаниях модели производится повторное обучение модели на следующей итерации и цикл повторяется. Для определения итерации с наилучшей моделью используется метрика BLEU.

При этом необходимо подчеркнуть, что при изучении проблем решения задач было выялено, что наборы данных попросту неоптимизированы для задач определения локации. Так, например, в наборе данных COCO существует довольно много слов, которые никогда не используются в качестве обозначения места действия, а потому являются вредными для обучения нейросетей. Например, в наборе данных присутствуют категории «sandwich», «sunglass». Обе категории практически не используются в качестве места действия текста. Поэтому обучать сеть необходимо было бы на наборе данных, полностью избавленном от подобных категорий.

Выводы.

Для проверки работы системы в качестве предобученных моделей для дальнецшего дообучения использовалить BERT-base-uncased, ROBERTA, CodeBERT. По результатам тестирования на реальных текстовых данных (для тестирования использовались тексты на английском языке) на всех предобученных сетях были получены высокие показатели точности для предложений, содержащий локацию в тексте (с небольшим преимуществом у CodeBERT). Показатели точности на таких предложениях варьируются от 74 до 82 процентов.

Однако для предложений, не содержащих в тексте самой локации, точность остается низкой (около 5 процентов). Для решения проблемы планируются сначала расширить набор данных за счёт предложений, сгенерированных с помощью GPT-3 ддя нужных локаций, а затем использовать синтетический набор данных для дообучения полученной модели и узнать на основе результатов, улучшится ли точность на предложениях без локаций и не уменишиться ли точность на предложениях с локацией. Если же условия не буду выполнятся для модели, то в таком случае планируется сгенерировать новую предобученную модель на основе генераторов, которая специализировалась именно на генерации выходных метод на основе входных предложений.

Мицковский Д.Ю. (автор)

Подпись

Фильченков А.А. (научный руководитель)

Подпись

Ефимова В.А. (научный консультант)

Подпись