УДК 004.85

ОБЗОР МЕТОДОВ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ПРИМЕРОВ ДЛЯ ОБУЧЕНИЯ ВОПРОСНО-ОТВЕТНЫХ СИСТЕМ

Ефимов П.В. (Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к.т.н., доцент Муромцев Д.И.
(Университет ИТМО, г. Санкт-Петербург)

В работе представлен обзор методов получения синтетических данных для обучения вопросно-ответных систем. Наиболее сложной задачей является генерация вопросов на основе отрывка текста и потенциального ответа. Большинство существующих методов были созданы для работы на английском языке и на искусственных данных. Поэтому в рамках работы наибольший интерес представляют методы, которые не требуют специфичных датасетов и инструментов и могут быть масштабированы на различные языки и домены.

Введение. Вопросно-ответные системы в отличие от информационного поиска позволяют быстрее удовлетворять информационные потребности пользователя возвращая краткий ответ вместо ранжированного списка документов. Подобные системы становятся частью Интернет-поиска, диалоговых систем, корпоративных хранилищ документов. Для того, чтобы обучить вопросно-ответную систему для определенного языка или области применения, необходим отдельный обучающий набор данных. Автоматически сгенерированные данные позволяют решить эту проблему. Такие данные можно использовать как для обучения с нуля, так и для расширения существующих наборов данных.

Основная часть. В качестве источника данных для обучения и/или оценки генератора вопросов часто используются датасеты, изначально предназначенные для вопросно-ответного поиска: SQuAD, Natural Questions, TriviaQA, MS MARCO. В качестве метрики качества генерации обычно используется метрика BLEU-4, пришедшая из машинного перевода. Также часто используются метрики METEOR и ROUGE.

Приемлемого качества вопросно-ответной системы можно добиться, обучившись на вопросах, полученных с помощью различных эвристик: перестановки слов в предложении с пропущенным ответом, вставки случайных вопросительных слов. Также такие данные можно использовать для расширения существующих датасетов, например, при межъязыковом переносе обучения.

Перспективным является использование современных порождающих моделей GPT на базе архитектуры Трансформер. Данные модели, предобученные на большом объёме данных, хранят в себе информацию о текстах и структуре языка. На основе нескольких обучающих примеров (few-shot learning) модель уже способна генерировать вопросы. Минусом данных моделей является то, что подобные моноязычные модели сейчас доступны для малого количества языков. Но текущие исследования показывают, что даже англоязычные модели способны генерировать текст на других языках за счет небольшого количества многоязычных данных попавших в корпус для предобучения. Также ведётся работа над получением многоязычных моделей GPT.

Для обучения генератора вопросов также используются модели, которые ранее применялись для машинного перевода (например, XLM, mT5).

Большое развитие получили методы “вопросно-ответного поиска без учителя”. В рамках данного подхода для генерации вопросов обучается модель машинного перевода, которая требует большой объём непараллельных данных (для генерации вопросов это несвязанные наборы предложений и вопросов). Последующие исследования данного подхода сейчас направлены на повышение разнообразия вопросов и снижения зависимости от текстов для поиска.

Выводы. Рассмотрено большое количество методов и подходов к автоматической генерации данных (вопросов) для обучения вопросно-ответных систем. Дальнейшее развитие данных подходов должно преследовать следующие цели:

1. применение данных подходов к языкам отличным от английского;
2. повышение качества и разнообразия получаемых вопросов;
3. снижение “похожести” между получаемыми вопросами и текстами для поиска ответа.