УДК 004.89

ПРИМЕНЕНИЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ГЕНЕРАЦИИ ВОПРОСОВ К ТЕКСТУ

Свищев А.Н. (Университет ИТМО),
Научный руководитель – к. ф.-м. н. Рыбин С.В.
(Университет ИТМО)

В работе рассматривается новая задача автоматической генерации вопросов в разговоре по заданному тексту. Предлагается воспроизвести и улучшить результат решения для англоязычных данных на русскоязычном переводном датасете.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 619423 «Разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе аргументационного подхода и глубокого машинного обучения»

Задача автоматической генерации вопросов в разговоре по заданному тексту (англ. conversational question generation, CQG) является важной для измерения способности машин вести разговор в вопрос-ответном стиле. Она может служить важным компонентом интеллектуальных диалоговых агентов или обучающих систем, задавая осмысленные и последовательные вопросы собеседникам для проверки их понимания определенной темы.

CQG задача заключается в том, чтобы на основе некоторого заданного текста и контекста беседы — вопрос-ответных пар на темы и факты из него, сгенерировать следующий вопрос. Она тесно связана с CQA (англ. conversational question answering) задачей, которая в свою очередь развивает проблему QA (англ. question answering). Однако, QA и QG задачах генерация ответа на вопрос к тексту или генерация вопроса к тексту являются единичным актом, в то время как в CQA и CQG требуется осуществлять связанную последовательность таких действий, учитывая контекст беседы и кореферентные связи в нем, также требуется глубокое понимание того, на какую информацию из текста следует обращать внимание для генерации последующего вопроса.

Основной проблемой при решении задачи генерации вопросов в разговоре по заданному тексту на русском языке является отсутствие датасетов для обучения моделей. Для QA и QG задачи существует SberSQuAD датасет, но его формат предполагает только наличие вопрос-ответных пар к фактам и текста, никак не связанных в естественную беседу. Для поиска решений и их валидации в CQA и CQG задачах совсем недавно был предложен CoQA датасет на английском языке.

В работе на первом этапе предлагается осуществить полуавтоматический перевод англоязычных данных на русский язык. Для оптимизации и упрощения процесса перевода разработано специальное приложение.

На втором этапе на переведенном наборе данных обучается глубокая нейронная сеть (Reinforced Dynamic Reasoning (ReDR) network). Предлагается обучить NMT (англ. neural machine translation, NMT) модель с копирующим механизмом внимания для генерации вопросов. Уже обученную модель также предлагается дообучать методом обучения с подкреплением, где в качестве вознаграждения выступает оценка качества ответа предобученной QA модели на сгенерированный вопрос. Это позволяет получать более разнообразные вопросы с высокой релевантностью темы вопроса тексту.

На третьем этапе предлагается расширить набор данных новыми автоматическими сгенерированными примерами. Для этого предобученные CQG и QA модели используются для последовательной генерации беседы (связанных вопрос-ответных пар) к заданному тексту. Последующий ручной отбор и доработка достаточно качественных примеров бесед существенно упрощает и удешевляет процедуру расширения набора данных.

Результатом ожидается получения обученной русскоязычной CQG модели интегрируемой в виртуального диалогового помощника для поддержки проведения дистанционного экзамена. Также побочным результатом работы планируется получить расширенный русскоязычный набор данных для изучения задач CQA и CQG.