УДК 004.855.5

РАЗРАБОТКА ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЫ ДЛЯ ДИАЛОГОВ НА IT ТЕМЫ НА ОСНОВЕ RUDIALOGPT

Самигулин Т. Р. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»), **Сафонова А. О.** (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»), **Хлюпина Ю. М.** (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»), **Демин О. Д.** (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Научный руководитель - доктор технических наук, доцент Басов О. О. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Аннотация. В докладе демонстрируется вопросно-ответная система для общения на IT-темы на основе русскоязычной DialoGPT - RuDialoGPT. Описываются этапы подготовки данных, а также дообучение модели RuDialoGPT.

Введение. В последнее время в области обработки естественного языка получило распространение GPT-3 - третье поколение алгоритма в данной области и одна из самых крупных и продвинутых языковых моделей на данный момент. Недавно ПАО Сбербанк представили русскоязычную версию данной модели ruGPT-3.

И хотя данные модели показывают высокие результаты в генерации осмысленного текста, было обнаружено, что они плохо подходят для вопросно-ответных систем. С этой целью корпорация Microsoft представил дообученную модель GPT-2 для диалогов - DialoGPT. Позже, вдохновившись модель DialoGPT, компания ICL Services представила русскоязычную модель DialoGPT - RuDialoGPT.

Данная модель показала хорошие результаты в общении с людьми и продемонстрировала умение следить за ходом диалогом на общие темы, но на специализированные темы, например IT, модель не всегда выдавала релевантные ответы. Поэтому, в данной работе предлагается способ для повышения качества ответов на диалоги, связанные с IT-тематикой.

Основная часть. В данном исследовании мы дообучили модель RuDialoGPT на основе собранных данных из открытых источников, чтобы повысить качество ответов на IT-тематики.

В связи с отсутствием готовых наборов данных, первым этапом стал сбор данных из открытых источников: Habr Q&A, Telegram IT чаты. Оба источника содержат вопросы, ответы, а также обсуждения IT-тем. Сбор данных начинался с исходного вопроса, например:“Какой язык программирования стоит выучить?”, после чего шла цепочка из 5 ответов на данный вопрос.

Далее данные были разделены на 2 части: тренировочная и тестовая. Процесс дообучения был запущен RuDialoGPT на тренировочной части. После этого, дообученная модель была проверена на тестовой выборке, а также проведены небольшие диалоговые сессии с моделью для человеческой оценки способности общения полученной модели на IT-темы.

Выводы. Полученная модель показала некоторые знания IT-темы, но она ещё далека от совершенства.

Для получения большего качества ответов нужен большой подготовленный набор данных. Возможным вариантов видится использования транскрипта IT сериалов и фильмов, так как они в явном виде содержат обсуждения на IT-темы.

Также следует добавить ранжировку сгенерированных ответов, чтобы добавить увеличение способности модели общаться с пользователем на связанные с IT темы.

Кроме того, в дальнейшем при обучении можно добавлять новые тематики для специализированных диалогов, например, финансовые технологии, геймдев и др. Это предоставит возможность управления тематикой генерируемых ответов в целях повышения их персонализации.

Самигулин Т. Р. (автор)	Подпись
Сафонова А.О. (автор)	Подпись
Хлюпина Ю.М. (автор)	Подпись
Демин О.Д. (автор)	Подпись
Басов О.О. (научный руководитель)	Подпись