

ПРИМЕНЕНИЕ НЕЙРОСЕТЕВОЙ МОДЕЛИ BLENDERBOT ДЛЯ ГЕНЕРАЦИИ РЕПЛИК НА РУССКОМ ЯЗЫКЕ

Апанасович К.С. (Университет ИТМО)

Научный руководитель – к.т.н. Махныткина О.В.

(Университет ИТМО)

В данной работе рассматриваются различные подходы к обучению нейросетевой модели BlenderBot для генерации реплик на русском языке.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №621296 «Разработка технологий для персонификации разговорного искусственного интеллекта».

Разработка разговорных агентов является очень востребованной во многих областях промышленности, а также в повседневной жизни. Чат-ботов принято разделять на две категории: closed-domain и open-domain. Closed-domain системы предназначены для решения конкретных заранее определенных задач. Open-domain системы не привязаны к какой-либо конкретной задаче или области знаний, а потому могут вести диалог с пользователем для любые темы. Однако такие агенты обладают рядом недостатков, один из которых – это непоследовательность в своих высказываниях.

Современные нейросетевые модели, такие как BlenderBot показывают хорошие результаты в задаче ведения диалога с пользователем. Однако, данная модель, обученная на англоязычных наборах данных, не приспособлена для генерации реплик на русском языке.

В данной работе была предпринята попытка адаптировать модель BlenderBot к ведению диалога на русском языке. Для обучения модели использовался русскоязычный набор данных Toloka Persona Chat Rus. Датасет содержит в себе 10013 диалогов, каждый из которых сопровождается описанием персон собеседников. Каждое описание представлено в виде 5 коротких предложений, например «Я рисую» или «Я живу за границей».

Также для адаптации модели к русскому языку будут использованы и протестированы следующие подходы:

- трансферное обучение (Transfer Learning);
- исправление ошибок в датасете с помощью готовых ресурсов, например DeepPavlov или Яндекс.Спеллер;
- аугментация данных.

Целью трансферного обучения является применение знаний, полученных из одной задачи, для решения другой целевой задачи. Реализация данного подхода заключается в том, чтобы вместо обучения всей нейросети сразу обучить лишь некоторые заранее определенные слои.

Датасет Toloka Persona Chat Rus содержит в диалогах значительное количество различных ошибок, которые могут негативно повлиять на обучение нейросетевой модели. Предполагается, что с помощью готовых ресурсов можно исправить значительную часть ошибок в диалогах и получить лучшие результаты после обучения.

Используемый набор данных является недостаточно большим для обучения модели с нуля, потому было решено использовать различные методы аугментации данных для получения большего набора данных.