УДК 004.89

ОПТИМИЗАЦИЯ ГИПЕРПАРАМЕТРОВ ПОИСКОВЫХ МОДЕЛЕЙ ДЛЯ ДИАЛОГОВЫХ ДАННЫХ

Посохов П.А. (Университет ИТМО)
Научный руководитель – к. т. н. Махныткина О.В.
(Университет ИТМО)

Введение. В данной работе проводится исследование ранжирующих поисковых моделей для задачи моделирования диалога на естественном языке. Представлено описание архитектуры ранжирующей модели, предложены целевые метрики и критерии сравнения, проведено исследование основных гиперпараметров: блоки кодирования, методы пулинга, функции сходства, функции потерь, и последующий их сравнительный анализ. Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 622282 «Разработка русскоязычного персонифицированного диалогового агента с динамической долгосрочной памятью».

Основная часть. Ранжирующие модели поиска – архитектура нейронных сетей, имеющая различные способы применения в рамках диалоговой системы, включая продуцирование ответов и использование в рамках систем генерации, основанной на знании. Ввиду этого исследование ранжирующих моделей, является актуальной задачей. Для проведения исследований был выбран набор текстовых диалоговых данных Toloka RuPersonaChat. Данный датасет включает 10013 диалогов, средняя продолжительность диалога – 16 реплик, средняя длина сообщения – 9 слов. В качестве базовой ранжирующей архитектуры была выбрана модель Bi-Encoder [1]. Данная архитектура предполагает ранжирование доступных кандидатов для рассматриваемого запроса в соответствии с их релевантностью от большей к меньшей. Запросы и кандидаты представляются в виде вектора произвольной размерности, который преобразуется в одномерный вектор при помощи пуллинга [2]. Это необходимо для ранжирования кандидатов на основе функции подобия векторов. Оптимизация модели производится на основе функции потерь от коэффициента подобия векторов.

Выводы. В данной работе были реализованы ранжирующие архитектуры с использованием BERT и T5-Encoder кодирующих блоков, обучаемые в синхронном и независимом режимах. Проведено сравнение различных методов пуллинга включая cls-pooling, mean-pooling и его модификации, stat-pooling и его модификации. Также была выбрана лучшая функция подобия векторов среди доступных: евклидова и манхэттенского расстояния, скалярного произведения и косинусной близости. Проведена реализация и анализ функций ошибки ранжирования: triplet loss, contrastive loss, cross-entropy loss. Произведена настройка лучших конфигураций, оптимизация моделей и сравнение полученных метрик качества.

Список использованных источников:

1. Humeau S. et al. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring //arXiv preprint arXiv:1905.01969. – 2019.
2. 2. Karpukhin V. et al. Dense passage retrieval for open-domain question answering //arXiv preprint arXiv:2004.04906. – 2020.