УДК 004.912

СРАВНИТЕЛЬНЫЙ АНАЛИЗ КОДИРУЮЩИХ МОДЕЛЕЙ ДЛЯ ZERO-SHOT ЗАДАЧ РАНЖИРОВАНИЯ

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

Введение. Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №623088 «Разработка русскоязычного персонифицированного эмоционального диалогового агента».

Тенденции последних десятилетий к накоплению текстовой неструктурированной или слабо структурированной информации на естественном языке обуславливают потребность в разработке методов их анализа. Рынок текстовой аналитики в 2024 году оценивается в 10.5 млрд долларов, и ожидается, что среднегодовой темп роста составит 39.90%. В список задач текстовой аналитики входят: поиск ответов на вопросы, моделирование диалога, поиск перефразирования, поиск по смысловым фрагментам и пр. Каждая из вышеприведенных задач требует использования собственной кодирующей модели и как следствие векторной базы кандидатов. Многозадачные модели могут разрешить вышеобозначенную проблему, а также повысить качество поиска в задачах с недостаточным количеством обучающих данных. Ввиду этого исследование кодирующих моделей, является актуальной задачей.

Основная часть. Для проведения исследований был отобран следующих список задач: поиск перефразирования, включающий 3 набора данных; поиск семантически близких текстов, включающий 3 набора данных; поиск ответов на вопросы, включающий 5 наборов данных; моделирование диалога, включающее 2 набора данных.

В качестве архитектуры ранжирующего поиска была выбрана модель Bi-Encoder. Данная архитектура предполагает ранжирование доступных кандидатов для рассматриваемого запроса в соответствии с их релевантностью от большей к меньшей. Кодирование запросов и кандидатов производится кодирующей моделью, которая переводит текстовую последовательность в виде токенов в вектор произвольной размерности, который затем должен быть преобразован в одномерный вектор при помощи пуллинг функции. Ранжирование кандидатов производится с использованием функции подобия векторов. Оптимизация модели производится на основе функции потерь от коэффициента подобия векторов.

Выводы. В данной работе исследуются предварительно обученные кодирующие модели типа BERT различных конфигураций, MPNet[1], ST-5[2]. Проводится сравнение качества с обучением и в режиме zero-shot и производительности данных кодирующих моделей в рамках ранжирующей Bi-Encoder архитектуры. Так же предлагается методика многозадачного предварительного обучения, повышающая качество поиска на большинстве рассматриваемых задач для кодирующей модели BERT tiny, обладающей лучшей производительностью поиска.

Список использованных источников:

1. Song K., Tan X., Qin T., Lu J., Liu T. Mpnet: Masked and permuted pre-training for language understanding //Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 16857-16867.

2. Zhang W., Xiong C., Stratos K., Overwijk A. Zhang W. et al. Improving Multitask Retrieval by Promoting Task Specialization //Transactions of the Association for Computational Linguistics. – 2023. – Т. 11. – С. 1201-1212.