УДК 004.89

ИССЛЕДОВАНИЕ ОГРАНИЧЕНИЙ ПРИ УМЕНЬШЕНИИ РАЗМЕРНОСТИ ВЕКТОРОВ В МОДЕЛЯХ ПОИСКА

Маслюхин С.М. (Университет ИТМО, ООО «ЦРТ-Инновации»)
Научный руководитель – д.т.н. Матвеев Ю.Н.
(Университет ИТМО)

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №622282 «Разработка русскоязычного персонифицированного диалогового агента с динамической долгосрочной памятью».

Введение. В настоящее время широкое применение получили нейросетевые модели поиска. Они используются как непосредственно в задачах поиска, так и в задачах ответа на вопросы и при моделирования диалога. Применение моделей поиска предполагает кодирование всех имеющихся текстов, по которым производится поиск, в вектора. Поисковая база в реальных системах может включать в себя сотни миллионов и даже миллиарды векторов, что приводит к существенным требованиям по объёмам оперативной памяти. Уменьшешение размерности векторов позволяет существенно сократить объём оперативной памяти, необходимой для хранения базы, однако следует учитывать ограничения, которые могут влиять на эффективность поиска по сжатым векторам [1]. Влияние различных факторов, влияющих на эффективность поиска, на данный момент мало изучено и остаётся актуальной задачей.

Основная часть. Для оценки ограничений при уменьшении размерности векторов в моделях поиска реализована модель поиска на основе данных перефразирования [2]. В состав данных входят различные датасеты перефразирования, представленные в открытом доступе, объединённые в единый набор данных и разбитые на обучающую, валидационную и тестовую выборку. Отдельно тестовая выборка разбита на подвыборки по длинне примеров после токенизации. В качестве модели выбрана rut5-small поскольку она обеспечивает высокое качестве решения целевой задачи и оптимальное время проведения экспериментов. Первая часть исследования посвящена оценке влияния длины текстов на эффективность поиска по векторам уменьшенного размера. Обученные с разной степенью сжатия выходных векторных представлений модели оценены на подвыборках с разной длинной тестовых примеров. По результатам проведённого эксперимента можно сделать вывод, что более длинные примеры требуют векторов большей размерности для сохранения эффективности поиска. Вторая часть исследования посвящена оценке влияния размеров поисковой базы на на эффективность поиска по векторам уменьшенного размера. Обученные с разной степенью сжатия выходных векторных представлений модели оценены на подвыборках с разным количеством примеров в поисковой базе.

Выводы. Результаты исследования позволили значительно сократить размер поисковой базы без заметных потерь в эффективности поиска.

Список использованных источников:

1. Raunak V. Simple and effective dimensionality reduction for word embeddings //arXiv preprint arXiv:1708.03629. – 2017.
2. Fenogenova A. Russian paraphrasers: Paraphrase with transformers //Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. – 2021. – С. 11-19.