УДК 004.8

РАЗРАБОТКА МЕТОДОВ ПОДГОТОВКИ ОБУЧАЮЩИХ ВЫБОРОК ДЛЯ МОДЕЛЕЙ СЕМАНТИЧЕСКОГО ПОИСКА

Лещенко С. Д.[1]
Научный руководитель – Маслюхин С. М.[2]
[1]Университет ИТМО, [2]ООО «ЦРТ-Инновации»
leshchsergei@gmail.com

Введение

Модели семантического поиска, основанные на языковых нейросетевых моделях и обучаемые с использованием контрастивного обучения, широко используются в задачах информационного поиска и рекомендательных системах. Качество таких моделей во многом зависит от стратегий формирования обучающих выборок. Особую роль при обучении играют сложные негативные примеры (hard negatives) [1], которые находятся близко к запросу в семантическом пространстве и позволяют формировать более информативный обучающий сигнал. Однако их автоматический отбор может приводить к появлению ложных негативов, ухудшающих процесс обучения моделей [2]. В связи с этим актуальной задачей является разработка методов подготовки обучающих выборок для моделей семантического поиска, ориентированных на эффективный отбор и использование сложных негативных примеров.

Основная часть

В работе рассматриваются методы формирования обучающих выборок для моделей семантического поиска в рамках подхода контрастивного обучения. Основное внимание уделяется стратегиям автоматического отбора сложных негативных примеров на основе оценки семантической близости, полученной языковой моделью [3].

Анализируются подходы, в которых в качестве источника негативных примеров используются результаты ранжирования документов или текстов различными языковыми моделями, что позволяет выявлять наиболее информативные, но трудные для различения примеры. Рассматриваются критерии отбора таких негативов и их влияние на процесс обучения и качество итоговых векторных представлений.

Дополнительно обсуждаются методы повышения качества обучающих выборок за счет исключения или ослабления влияния некорректных и шумных примеров, возникающих в процессе автоматического отбора данных. Предлагается обобщенный пайплайн подготовки обучающих данных, включающий этапы генерации, отбора и последующей валидации сложных негативных примеров. Эффективность подходов оценивается в задаче семантического поиска с использованием стандартных метрик качества ранжирования [4].

Выводы

В рамках работы показано, что использование сложных негативных примеров при формировании обучающих выборок позволяет добиться повышения качества моделей семантического поиска. Разработка методов их автоматического отбора и контроля качества позволяет получить более устойчивые и информативные обучающие данные. Предложенные подходы могут быть использованы при построении и оптимизации практических систем семантического поиска и интегрированы в существующие пайплайны обучения нейросетевых моделей.

Литература

1. Moreira G. S. P. et al. NV-Retriever: Improving text embedding models with effective hard-negative mining // arXiv preprint arXiv:2407.15831. – 2024.
2. Mitigating False Negatives in Multiple Negatives Ranking Loss for Retriever Training [Электронный ресурс] // Hugging Face. – URL: https://huggingface.co/blog/dragonkue/mitigating-false-negatives-in-retriever-training (дата обращения: 03.02.2026).
3. Lee C. et al. Nv-embed: Improved techniques for training llms as generalist embedding models // arXiv preprint arXiv:2405.17428. – 2024.
4. Demystifying NDCG [Электронный ресурс] // Towards Data Science. – URL: https://towardsdatascience.com/demystifying-ndcg-bee3be58cfe0 (дата обращения: 03.02.2026).