УДК 004.89

ИССЛЕДОВАНИЕ МЕТОДОВ ВЫБОРА РЕЛЕВАНТНЫХ ЭЛЕМЕНТОВ В ДИАЛОГОВЫХ СИСТЕМАХ С ДОСТУПОМ К НЕСТРУКТУРИРОВАННЫМ ДАННЫМ

Маслюхин С.М. (ООО «ЦРТ-Инновации»)
Научный руководитель – доктор технических наук, Матвеев Ю.Н.
(ИТМО)

Введение. Диалоговые системы, способные взаимодействовать с пользователем на естественном языке, становятся все более востребованными. Ключевой задачей для таких систем, особенно при работе с большими объемами неструктурированных данных (отзывы, FAQ), является эффективный выбор релевантных информационных элементов. От точности этого выбора напрямую зависит качество ответа системы и, как следствие, удовлетворенность пользователя. В данном исследовании рассматривается задача выбора релевантных фрагментов текста из неструктурированных источников (отзывы клиентов ресторанов и отелей, FAQ-документы) для формирования ответа диалоговой системы на запрос пользователя. Цель исследования - анализ и сравнение различных методов выбора релевантных элементов для повышения качества работы диалоговой системы. Актуальность работы обусловлена ростом объема доступной неструктурированной информации и необходимостью автоматизации её обработки.

Основная часть. В рамках данного исследования рассматривается задача поиска релевантных элементов в контексте диалоговых систем, оперирующих неструктурированными данными. В качестве источников данных используются отзывы клиентов ресторанов и отелей, а также FAQ-документы, представляющие собой неструктурированный текст. Для решения задачи извлечения релевантных фрагментов, способных служить основой для ответа на вопрос пользователя, были протестированы различные подходы, основанные на оценке семантической близости между запросом и элементами поисковой выдачи.

Центральным элементом исследования является использование модели NV-Embed-v2 [1] для получения векторных представлений (эмбеддингов) запросов и текстовых фрагментов. Эти эмбеддинги используются для вычисления меры близости, на основе которой принимается решение о релевантности. В качестве базового подхода (baseline) используется оценка F1, рассчитываемая при выборе наилучшего глобального порога для всех запросов. Эксперименты проводились на данных, предоставленных в рамках конкурса DSTC11 Track 5 [2]. Базовый подход продемонстрировал результат F1 = 0.165 при пороге 11.

Далее были исследованы два метода, направленных на улучшение результатов baseline: AttnCut [3] и CosineAdapter [4]. AttnCut, использующий механизм внимания для адаптивного выбора порога, позволил улучшить базовый результат до F1 = 0.178. Это свидетельствует о том, что адаптивный подход к определению порога релевантности может быть более эффективным, чем использование единого глобального порога.

Подход CosineAdapter, который оперирует порогом на уровне значений косинусной близости, показал незначительное улучшение по сравнению с baseline (F1 = 0.168). Этот результат объясняется тем, что CosineAdapter работает с порогом на уровне значений скоров, а для данного датасета (DSTC11 Track 5) порог по скорам изначально демонстрирует более низкую эффективность.

Выводы. Проведенное исследование различных методов выбора релевантных элементов в диалоговых системах, работающих с неструктурированными данными, показало, что адаптивные методы, такие как AttnCut, могут превосходить базовый подход, основанный на глобальном пороге. Использование модели NV-Embed-v2 для получения векторных представлений запросов и текстовых фрагментов обеспечивает основу для эффективного

сравнения семантической близости. Результаты, полученные на данных конкурса DSTC11 Track 5, продемонстрировали, что AttnCut улучшает F1-меру с 0.165 (baseline) до 0.178. CosineAdapter показал незначительное улучшение. Полученные результаты подчеркивают важность дальнейших исследований в области адаптивных методов выбора релевантных элементов для повышения качества работы диалоговых систем с доступом к неструктурированным данным.

Список использованных источников:

1. Lee C. et al. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models //arXiv preprint arXiv:2405.17428. – 2024.
2. Zhao C. et al. " What do others think?": Task-Oriented Conversational Modeling with Subjective Knowledge //arXiv preprint arXiv:2305.12091. – 2023.
3. Wu C. et al. Learning to truncate ranked lists for information retrieval //Proceedings of the AAAI Conference on Artificial Intelligence. – 2021. – Т. 35. – №. 5. – С. 4453-4461.
4. Rossi N. et al. Relevance filtering for embedding-based retrieval //Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. – 2024. – С. 4828-4835.