

Задание 1.4. ИСР

«Рецензирование научной статьи (по проблеме, исследуемой в магистерской диссертации). Написать рецензию на статью»

Автор:

Воложанин Владислав Олегович

1. Библиографическое описание рецензируемой статьи

Mulla, N., Gharpure, P. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 2023, Vol. 12, pp. 1–32. Опубликована онлайн 30 January 2023. DOI: 10.1007/s13748-023-00295-9.

2. Актуальность статьи и связь с темой магистерской диссертации

Рецензируемая статья является обзорной работой по автоматической генерации вопросов (Automatic Question Generation, AQG) направлению, напрямую связанному с темой магистерской диссертации «Веб-приложение для микрообучения сотрудников с интерактивными заданиями, автоматической генерацией тестов и аналитикой». AQG в прикладном корпоративном контексте выступает ключевой технологической основой для масштабируемой подготовки тестов и интерактивных проверок без постоянного ручного участия экспертов и методистов.

Особая актуальность статьи для диссертации состоит в том, что авторы рассматривают не только методы генерации, но и наборы данных и методы оценки качества то есть те элементы, от которых зависит практическая пригодность автогенерации тестов в реальном веб-приложении (валидность вопросов, их «естественность», уместность и измеримость качества).

3. Цель и задачи статьи (по содержанию аннотации и описанию работы)

Авторы заявляют, что предоставляют обзор прогресса исследований в области автоматической генерации вопросов и формируют классификацию систем QG по трём крупным сценариям использования:

1. standalone question generation (генерация вопросов по тексту/контенту),
2. visual question generation (вопросы по изображениям/видео),
3. conversational question generation (вопросы в диалоговых системах).

В качестве заявленных вкладов (contributions) авторы выделяют: (а) обзор методологий, (б) перечень/обзор датасетов для AQG, (в) обсуждение вызовов и приложений.

Также в аннотации описан принцип отбора публикаций: использовались ключевые фразы для поиска, далее применялись критерии включения, ориентированные на качество (работы из «repute» журналов/конференций, наличие экспериментальной проверки на бенчмарках, сравнения с существующими решениями и т.п.).

4. Краткая характеристика содержания и логики изложения

Работа имеет обзорный характер и строится вокруг типичного для survey-статьей каркаса:

- постановка задачи QG и обсуждение типов вопросов;
- классификация подходов по сценариям применения (standalone/visual/conversational);
- обзор доступных датасетов;
- обзор подходов к оценке качества вопросов (автоматическая и человеко-ориентированная оценка);
- перечисление практических применений и исследовательских проблем/вызовов.

Отдельно важно, что авторы подчёркивают наличие двух больших классов оценивания качества: *automatic evaluation* и *human-based evaluation*, а также отмечают, что широко используемые метрики автоматической оценки (например, BLEU/ROUGE/METEOR) не всегда отражают «естественность» и смысловую состоятельность вопроса, из-за чего сохраняется потребность в более содержательных метриках качества или комбинированных схемах оценки.

5. Научная и практическая значимость

Научная значимость статьи состоит в систематизации области: наличие единой классификации use-case'ов (standalone/visual/conversational) упрощает анализ литературы и помогает корректно позиционировать исследования и инженерные решения. Это особенно важно для диссертации, где требуется обоснованно выбрать «часть пространства задач», соответствующую домену корпоративного обучения (в первую очередь, standalone QG по инструкциям/регламентам/микроконтенту).

Практическая значимость для разработки веб-приложения заключается в том, что статья связывает три критических компонента внедрения автогенерации тестов:

1. методы генерации,
2. данные для обучения/валидации,
3. метрики/оценивание качества,
и тем самым помогает сформировать требования к «контуру качества» в системе (редактирование, фильтрация, рейтинг вопросов, сбор статистики качества).

6. Достоинства статьи

1. Широкий охват и структурирование области: рассмотрение сразу трёх сценариев (standalone, visual, conversational) делает обзор «карты местности» для исследователя и разработчика.
2. Фокус на оценивании качества: в реальных образовательных/корпоративных продуктах именно качество вопроса определяет пользу, а не сам факт генерации; акцент на проблеме метрик и необходимости более содержательной оценки делает обзор прикладно полезным.
3. Методологически осмысленный отбор работ (по заявлению авторов): наличие критериев включения/исключения повышает доверие к обзорной части и снижает риск «случайной библиографии».
4. Актуальные формальные сведения о публикации (для корректного цитирования в диссертации): указаны даты получения/принятия/публикации, выходные данные, DOI.

7. Замечания и дискуссионные моменты (критическая оценка)

1. Общая (недоменно-ориентированная) рамка обзора. Статья позиционируется как широкий обзор AQG в целом. Для корпоративного обучения часто требуются дополнительные критерии качества, которые в обзорных AQG-работах обычно раскрыты ограниченно: привязка вопроса к конкретной цели обучения/компетенции, соответствие корпоративным регламентам, трассируемость «вопрос → источник → правило», управляемая

сложность, качество дистракторов для MCQ и т.п. Это не столько недостаток, сколько ограничение «универсального обзора» при переносе в доменную задачу.

2. Проблема метрик качества остаётся открытой. Авторы фиксируют, что стандартизованные метрики генерации текста (BLEU/ROUGE/METEOR и др.) ограниченно отражают педагогическую ценность и «естественность» вопросов, а потребность в более релевантных метриках сохраняется. Для диссертационного проекта это означает, что одной только «оценки по BLEU» будет недостаточно; потребуется либо экспертная валидация, либо эксплуатационная аналитика качества вопроса по данным ответов пользователей.
3. Риск разрыва между исследовательскими датасетами и корпоративными данными. В аннотации подчёркивается важность датасетов и бенчмарков, однако корпоративные материалы (инструкции, политики, внутренние базы знаний) часто отличаются стилем и структурой от публичных датасетов. Следовательно, практическое внедрение в микрообучении потребует доменной адаптации, собственных наборов данных/правил контроля качества и осторожной интерпретации результатов обзора при выборе архитектуры.

8. Вывод и рекомендация (итог рецензии)

Статья Mulla & Gharpure представляет собой качественный обзор AQG, полезный как теоретическая и методическая основа для разделов магистерской диссертации, связанных с автоматической генерацией тестов: классификацией подходов, обзором датасетов и обсуждением оценки качества вопросов.

С учётом темы диссертации статью целесообразно использовать:

- в главе 1 (обзор литературы) как источник, систематизирующий AQG и подчёркивающий проблемы качества/оценки;
- при формировании требований к модулю генерации тестов: обязательность контура валидации, гибридная оценка (автоматическая + человеческая), хранение метаданных и источников;

- как аргумент, что “автогенерация” в образовательном продукте должна сопровождаться измерением качества (в т.ч. через аналитику результатов прохождения).

Рекомендация: статья может быть рекомендована к использованию в магистерской диссертации как базовый обзор по AQG, при обязательном дополнении доменно-ориентированными источниками по образовательной валидности вопросов, генерации MCQ-дистракторов и эксплуатационной аналитике качества заданий в цифровом обучении.