УДК 004.81

АВТОМАТИЧЕСКОЕ ДОБАВЛЕНИЕ SEO-МЕТАДАННЫХ В НОВОСТНЫЕ СТАТЬИ С ИСПОЛЬЗОВАНИЕМ QWEN-CODER

Х. Салем[1] [0000-0002-9143-5231], **А. С. Тощев**[2] [0000-0003-4424-6822]

[1]Университет Иннополис, г. Иннополис, Россия
[2]Казанский (Приволжский) федеральный университет, г. Казань, Россия

[1]h.salem@innopolis.ru, [2]atoschev@kpfu.ru

Аннотация

Обобщен ранее разработанный конвейер обогащения новостных статей структурированными метаданными и представлена его обновленная конфигурация, в которой GPT-3 (Generative Pre-trained Transformer 3) – языковая модель от компании OpenAI – заменен на открытую модель Qwen-Coder. Новая версия, как и ранее, использует набор из 400 страниц, отобранных через Google News, и остается совместимой с Google Rich Results Test. Эксперименты показали, что качество, сопоставимое с GPT-3, достижимо при локальном запуске на типовом офисном настольном компьютере (CPU, без GPU). Установлено, что замена, указанная выше, снижает зависимость от платных облачных сервисов и обеспечивает более высокую производительность по сравнению с GPT-версией; дана оценка сходства результатов обогащения для Qwen-Coder относительно базовой реализации на GPT-3. Предложенные инструменты снижают порог внедрения семантической разметки и расширяют ее практическое применение, в том числе в цифровой журналистике.

Ключевые слова: *семантическая паутина, майнинг шаблонов, Qwen-Coder, новостные веб-страницы, читабельность, структурированные данные.*

ВВЕДЕНИЕ

Семантическая разметка веб-страниц позволяет представлять их содержание и метаданные в машиночитаемом виде, благодаря чему программные системы могут корректнее интерпретировать материалы и формировать расши-

ренные элементы поисковой выдачи [1–2]. В новостном сегменте такие механизмы особенно важны: агрегаторы и поисковые сервисы (включая Google News) используют структурированные метаданные в формате JSON-LD для понимания типа материала, даты публикации, авторства и других характеристик [3–4]. Однако значительная часть издателей продолжает публиковать страницы в виде «чистого» HTML без структурированной разметки, что ограничивает потенциал поисковой видимости и снижает качество представления материалов в выдаче.

Ранее в работе [5] был предложен пятиэтапный конвейер обогащения новостных страниц: он собирает статьи, извлекает и очищает основной текст и формирует корректные объекты JSON-LD, описывающие метаданные страницы. В исходной реализации ключевая операция очистки и нормализации текста выполнялась с использованием GPT-3 (Generative Pre-trained Transformer 3) – крупной языковой модели третьего поколения семейства GPT от компании OpenAI [6].

В настоящей статье рассмотрена обновленная конфигурация этого конвейера, в которой вместо GPT-3 применена Qwen-Coder – открытая языковая модель семейства Qwen, ориентированная на задачи программирования и обработки структурированных форматов (в том числе разметки и сериализаций данных) [7]. Такая замена делает решение менее зависимым от облачных платных интерфейсов, уменьшает риски, связанные с внешними ограничениями и нестабильными задержками, и позволяет использовать локальный запуск в типовой корпоративной инфраструктуре. Это важно для организаций с ограниченными ресурсами и для команд, которым требуется масштабно формировать структурированные метаданные при отсутствии специализированной экспертизы в машинном обучении [8–9]. Дополнительно переход на открытую модель поддерживает курс отрасли информационных технологий на расширение доступности языковых моделей вне подписных сервисов [7].

В статье представлены обзор литературы, описание методики и результаты сравнения полученных результатов с GPT-версией по показателям качества и производительности. Отдельно рассмотрены вопросы эксплуатации и масштабирования решения в условиях эксплуатации.

ОБЗОР ЛИТЕРАТУРЫ

Как известно, семантический веб – это подход, при котором сведения на веб-страницах описывают так, чтобы их могли однозначно интерпретировать не только люди, но и программы. Для этого применяют формальные модели представления знаний и стандартные форматы описания сущностей и их связей [3, 8, 10]. В этой области широко используются RDF (модель представления фактов в виде «субъект – связь – объект») и OWL (язык для описания онтологий, то есть набора понятий предметной области и отношений между ними) [3, 10]. В прикладных задачах веб-публикации чаще применяют JSON-LD (формат добавления «связанных данных» в виде JSON), а также альтернативную разметку RDFa в составе HTML [4, 11–12].

Для новостных сайтов структурированные метаданные важны тем, что поисковые системы могут использовать их для более точного понимания материала (тип публикации, автор, дата, рубрика, изображение и т. п.) и формирования расширенных элементов выдачи [4, 11, 12–13]. Отдельные работы показали, что корректная поисковая оптимизация (SEO – набор приемов, повышающих видимость страниц в поиске) влияет на результативность продвижения и может быть связана с бизнес-показателями [9, 14–16]. При этом поведение пользователей в поисковой выдаче (SERP – страница результатов поиска) изучается как самостоятельная тема; такие исследования показывают, какие элементы выдачи привлекают внимание и как изменяются сценарии просмотра [17].

Практическая проблема заключается в том, что многие сайты остаются «слабоструктурированными»: они публикуют корректный HTML, но не добавляют формализованные метаданные или делают это непоследовательно. Особенно заметно это у агрегаторах новостей, где материалы поступают из большого числа доменов с различными шаблонами страниц; на примере Google News в [18] подробно проанализированы эффекты нормализации и различий между источниками. Поэтому в реальных системах часто используют комбинированный подход: часть метаданных извлекают по структуре страницы, а часть – по текстовому содержанию.

Отдельное направление работ посвящено извлечению основного содержимого с страниц (заголовка и «тела» новости) из HTML-документов. Такие методы используют DOM (Document Object Model – древовидное представление элементов HTML-страницы), визуальные признаки и устойчивые шаблоны расположения блоков, чтобы отделять основной текст от меню, рекламы и служебных элементов [19–24]. Эти исследования важны для нашей задачи, поскольку качество структурированных метаданных напрямую зависит от качества выделения основного контента.

Кроме того, в ряде работ особо подчеркнуто, что автоматическая разметка требует регулярного сопровождения знаний предметной области: словарей, правил, классов сущностей и их атрибутов. Без такой поддержки система постепенно теряет качество из-за изменений в доменах, шаблонах страниц и требованиях поисковых платформ [10]. В прикладных сценариях это означает необходимость контроля качества и периодической актуализации правил формирования метаданных, а также проверки результата внешними средствами валидации, например Google Rich Results Test [13].

МЕТОДОЛОГИЯ

Мы используем исходный корпус из 1100 статей на английском и арабском языках, выбранных из 18 источников через Google News – новостного агрегатора, который объединяет материалы различных изданий [18]. Обновленный конвейер, как и ранее, выполняет следующую последовательность шагов: (1) загрузку веб-страниц; (2) удаление имеющейся разметки JSON-LD; (3) извлечение признаков из DOM; (4) формирование структурированных метаданных с использованием языковой модели и (5) проверку результата. В новой версии этап генерации выполняется моделью Qwen-Coder [7].

Для генерации метаданных использованы ключевые поля страницы: заголовок, основной текст, основное изображение и адрес страницы (URL). Эти поля формируют текстовый запрос к языковой модели, который передается в локально запущенное приложение. На выходе формируется блок JSON-LD, совместимый с общепринятыми схемами описания контента (schema.org). Полученные результаты сохраняются в кэш и проходят проверку с помощью Google Rich

Results Test [13]. Дополнительно проводится количественная проверка сохранения смысла: вычисляется расхождение Дженсена–Шеннона [25–28] между исходным и восстановленным представлениями содержимого, что позволяет контролировать семантическую эквивалентность.

Переход на Qwen-Coder потребовал адаптации развертывания под процессорный режим (CPU, без использования графического ускорителя). Для этого модель была упакована в облегченную среду выполнения и использовала квантованные веса (компактное представление параметров модели для экономии памяти), а также пакетную обработку запросов с учетом ограничений оперативной памяти. Мы сохранили те же шаблоны запросов, что применялись в конфигурации с GPT-3, чтобы остальные компоненты конвейера (проверка, хранение и обработка результатов) работали без изменений. Такой подход уменьшил риск ошибок при переходе на Qwen-Coder и позволил измерить вклад именно замены модели.

Для сопровождения процесса были добавлены показатели контроля работы приложения: время обработки запросов, доля обращений, обслуженных из кэша (показывает эффективность повторного использования результатов), а также значение расхождения Дженсена–Шеннона. Эти показатели были использованы для оперативного выявления проблем, связанных с входными данными или ограничениями аппаратной платформы.

Нагрузочное тестирование (проверка поведения системы при росте объема обработки) выполнялось в виде повторяющихся пакетных запусков в течение одного дня с постепенным увеличением масштаба: ежедневные серии по 20 страниц, еженедельные обновления по 200 страниц и архивные догрузки по 400 страниц. Каждый запуск сопровождался одинаковыми процедурами проверки результата, что позволило оценить устойчивость работы при длительной нагрузке и при разных сценариях поступления данных. В результате экспериментов были получены показатели производительности и вариативности времени обработки, которые далее использовались в сравнительном анализе.

ВОЗМОЖНОСТИ QWEN3-CODER ПО СРАВНЕНИЮ С GPT-3

Qwen3-Coder нельзя рассматривать как простую замену «один к одному» по отношению к GPT-3. Во-первых, эта модель поддерживает расширенный контекст, то есть может обрабатывать значительно больший объем входного текста в одном запросе (до 256 тыс. токенов) [7]. Во-вторых, по заявлению разработчиков, она ориентирована на широкий набор языков программирования и форматов разметки, а также поддерживает режим работы с внешними инструментами: модель может не только генерировать текст, но и выполнять последовательность действий, вызывая подключенные функции по мере необходимости [29].

Эти возможности важны для нашего конвейера: они позволяют обрабатывать более крупные пакеты статей без жесткого упрощения входного запроса и сохранять корректную работу со структурированными форматами, включая JSON LD. Для версии на GPT-3 требовалось сокращать входные данные, чтобы уложиться в ограничение на длину запроса (порядка 4–8 тыс. токенов). В результате сочетание локального развертывания, расширенного контекста и ориентации на структурированные форматы стало ключевой мотивацией миграции на Qwen-Coder [7].

РЕЗУЛЬТАТЫ

На подмножестве из 400 страниц применение Qwen-Coder обеспечило в среднем 93% сходства между обновленным и исходным текстами статьи. При этом итоговые страницы сохраняли корректное отображение материала, включая встроенные изображения и графические элементы. Локальный запуск увеличил среднее время обработки примерно на 1.5 с по сравнению с вариантами, основанными на GPT-3, однако такой скорости оказалось достаточно для выполнения ночных пакетных обработок.

Отказ от использования платных облачных сервисов позволил снизить прогнозируемые ежемесячные расходы на 68% с учетом затрат на оборудование, лицензирование и поддержку [8]. Кроме того, выполнение всех этапов обработки внутри локальной инфраструктуры упростило соблюдение внутренних требований организаций к защите и контролю данных.

Помимо усредненных показателей были рассмотрены результаты в разрезе источника публикации, языка и длины статьи. Во всех сегментах сохранялись высокие значения сходства. Более заметная вариативность наблюдалась у длинных материалов с большим числом встроенных медиа-элементов и ссылок. Ручная проверка показала, что отклонения чаще связаны с неоднородной структурой HTML-страниц разных сайтов, а не с ошибками модели; это указывает на потенциал дальнейшего улучшения правил извлечения основного содержания.

По результатам применения подхода было отмечено ускорение диагностики и устранения ошибок валидации, поскольку журналы событий и промежуточные файлы формируются и сохраняются внутри локальной инфраструктуры. Это сделало работу конвейера более прозрачной и снизило потребность во внешней поддержке, что повысило общий экономический эффект.

В табл. 1 представлены количественные показатели сравнения GPT-3 и Qwen-Coder для трех размеров пакетов. Для каждого сценария приведены среднее время выполнения и стандартное отклонение по пяти повторениям. Хотя среднее время для наименьшего пакета сопоставимо, Qwen-Coder демонстрирует значительно меньшую дисперсию, что делает ночные операции более предсказуемыми. На рис. 1 наглядно видна разница времени обработки между GPT-3 и Qwen-Coder. На рис. 2 показано сравнение дисперсии, где более стабильное время выполнения показывает Qwen-Coder.

Табл. 1. Сравнение времени обработки для GPT-3 и Qwen-Coder при различных размерах пакетов. Меньшие значения означают более быстрые или более стабильные прогоны.

Размер пакета (страниц)	Среднее время, мин (GPT-3)	Среднее время, мин (Qwen-Coder)	Стандартное отклонение, мин (GPT-3)	Стандартное отклонение, мин (Qwen-Coder)
20	18.4	17.9	4.2	1.6
200	162.7	149.3	21.5	6.8
400	339.5	301.8	48.2	11.4

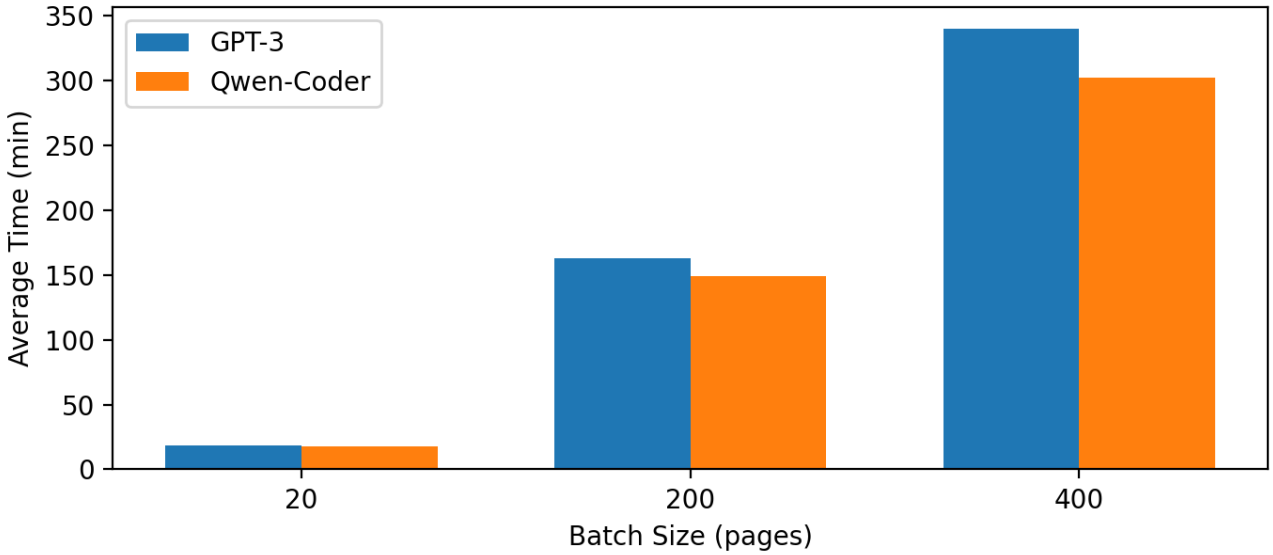

Рис. 1. Среднее время обработки для GPT-3 и Qwen-Coder при различных размерах пакетов.

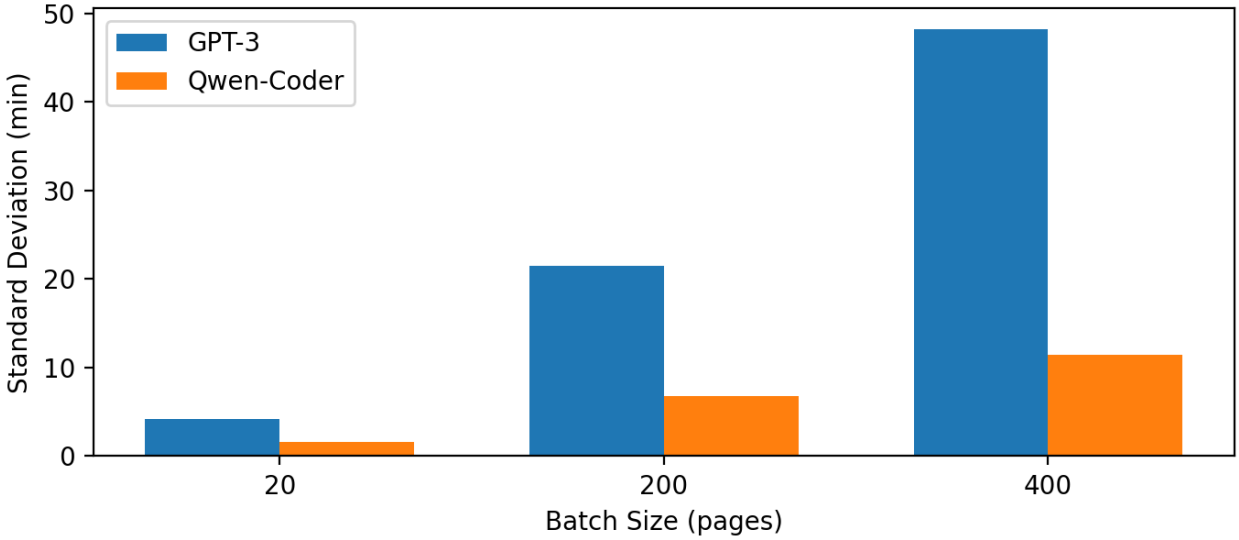

Рис. 2. Сравнение дисперсии: более стабильное время выполнения у Qwen-Coder.

Более подробный анализ данных мониторинга показал, что при локальном запуске Qwen-Coder реже возникают задержки, характерные для сетевых обращений: повторные попытки соединения и ограничения скорости со стороны удаленного сервиса, с которыми мы сталкивались в конфигурации на GPT-3. Даже при сопоставимой средней скорости обработки исчезновение редких, но

длительных задержек делает время выполнения более стабильным и упрощает планирование ночных и еженедельных пакетных запусков. Такая стабильность особенно важна для организаций, которым нужно обновлять структурированные подборки материалов до начала утреннего регионального новостного цикла.

ОГРАНИЧЕНИЯ И ПЛАНИРУЕМАЯ РАБОТА

Нами был использован корпус из 400 страниц, отобранных через Google News, который собирает публикации различных СМИ и предоставляет единый доступ к ним. Поскольку выборка ограничена материалами, попавшими в этот агрегатор, результаты могут не полностью переноситься на другие домены и типы сайтов (например, корпоративные порталы, блоги или специализированные площадки) с иной структурой страниц и правилами разметки. Мы сосредоточились на замене модели и сравнении производительности, оставив измерение долгосрочных SEO-эффектов, а также анализ потребления системой мощности для будущих исследований.

В дальнейшем планируется расширить набор данных и исследовать мультиязычные промпты. Мы также намерены экспериментировать с GPU-ускорением и измерять SEO-эффект за длительное время использования. Дополнительно включение показателей энергопотребления и доступности в систему мониторинга позволит более полно описать эксплуатационные компромиссы.

ЭКСПЛУАТАЦИОННЫЕ АСПЕКТЫ

Использование локально развернутой языковой модели предъявляет повышенные требования к эксплуатации и предполагает заранее формализованные процедуры сопровождения. Для обеспечения устойчивости при длительной работе конвейера организован регулярный контроль температуры процессора, загрузки оперативной памяти и состояния дисковой подсистемы. При приближении показателей к установленным порогам система автоматически формирует уведомления, что позволяет выполнять профилактические действия до возникновения отказов.

С учетом возможного роста вычислительной нагрузки разработан план обеспечения вычислительными ресурсами. Он предусматривает масштабирование за счет добавления вычислительных узлов, а также применение ускорения на графических процессорах в периоды пикового спроса (например, во время выборов или крупных спортивных событий). Соответствующие сценарии предварительно отрабатываются в контролируемой среде до перевода в промышленную эксплуатацию для подтверждения производительности и оценки затрат.

Отдельное внимание было уделено управлению программными зависимостями и восстановлению. Артефакты модели и конфигурационные файлы продублированы во внутреннем хранилище, что обеспечивает быстрое восстановление при отказе локального диска. Процедуры резервирования дополнительно тестировались в учебных испытаниях с имитацией сбоев инфраструктуры.

ЗАКЛЮЧЕНИЕ

Замена GPT-3 на Qwen-Coder позволила сохранить точность алгоритма добавления SEO-метаданных и повысить его производительность при работе на менее мощном оборудовании. Такая замена обеспечила существенную экономию, упростила развертывание и сохранила совместимость с алгоритмом, работающим с GPT-3. В дальнейшем планируется исследовать многоязычные шаблоны запросов и адаптивные механизмы кэширования с сохранением подхода локального запуска, который делает решение привлекательным для организаций с ограниченным бюджетом.

В более широком контексте результаты работы показали, что открытые языковые модели могут применяться в сценариях промышленной эксплуатации, которые ранее опирались на закрытые облачные сервисы. Представляя практический опыт внедрения и результаты оценки, мы рассчитываем поддержать дальнейшие исследования и разработки доступных инструментов, упрощающих применение практик семантической разметки в журналистике.

В дальнейшем планируется реализовать полуавтоматическую настройку запросов к модели, расширить языковое покрытие за пределы английского и арабского языков, а также углубить интеграцию с корпоративными системами управления контентом. Каждое изменение будет оцениваться с использованием

тех же метрик, что и в настоящей статье, чтобы повышение функциональности не снижало надежность и прозрачность результатов.

СПИСОК ЛИТЕРАТУРЫ

1. *Bashir F., Warraich N.F.* Systematic literature review of Semantic Web for distance learning // Interactive Learning Environments. 2020. Vol. 31. P. 527–543.

2. *Breit A., Waltersdorfer L., Ekaputra F.J., Sabou M., Ekelhart A., Iana A., Paulheim H., Portisch J., Revenko A., Teije A.T., et al.* Combining Machine Learning and Semantic Web: A Systematic Mapping Study // ACM Computing Surveys. 2023. Vol. 55. Art. 313.

3. *Yu L.* Introduction to the Semantic Web and Semantic Web Services. Boca Raton, FL, USA: Chapman and Hall/CRC, 2007.

4. *Sporny M., Longley D., Kellogg G., Lanthaler M., Lindström N.* JSON-LD 1.1: W3C Recommendation. 2020.

5. *Salem H., Salloum H., Orabi O., Sabbagh K., Mazzara M.* Enhancing News Articles: Automatic SEO Linked Data Injection for Semantic Web Integration // Applied Sciences. 2025. Vol. 15. Art. 1262. https://doi.org/10.3390/app15031262

6. OpenAI. GPT-3 powers the next generation of apps. 2021.
URL: https://openai.com/index/gpt-3-apps/ (дата обращения: 16.01.2026)

7. *Hui B., Yang J., Cui Z. et al.* Qwen2.5-Coder Technical Report // arXiv. 2024. arXiv:2409.12186. URL: https://arxiv.org/abs/2409.12186 (дата обращения: 10.01.2026).

8. *Shadbolt N., Berners-Lee T., Hall W.* The Semantic Web Revisited // IEEE Intelligent Systems. 2006. Vol. 21. P. 96–101.

9. *Poturak M., Keco D., Tutnic E.* Influence of search engine optimization (SEO) on business performance: Case study of private university in Sarajevo // International Journal of Research in Business and Social Science. 2022. Vol. 11. P. 59–68.

10. *Chandrasekaran B., Josephson J.R., Benjamins V.R.* What are ontologies, and why do we need them? // IEEE Intelligent Systems and Applications. 1999. Vol. 14. P. 20–26.

11. *Sporny M., Longley D., Kellogg G., Lanthaler M., Lindström N.* JSON-LD 1.0: W3C Recommendation. 2014.

12. *Adida B., Birbeck M., McCarron S., Pemberton S.* RDFa in XHTML: Syntax and processing: W3C Recommendation. 2008.

13. Rich Results Test. URL: https://search.google.com/test/rich-results (дата обращения: 08.10.2024).

14. *Iqbal M., Khalid M.N., Manzoor A.A., Malik M., Shaikh N.A.* Search Engine Optimization (SEO): A Study of important key factors in achieving a better Search Engine Result Page (SERP) Position // Sukkur IBA Journal of Computing and Mathematical Sciences. 2022. Vol. 6. P. 1–15.

15. *Alfiana F., Khofifah N., Ramadhan T., Septiani N., Wahyuningsih W., Azizah N.N., Ramadhona N.* Apply the Search Engine Optimization (SEO) Method to determine Website Ranking on Search Engines // International Journal of Cyber Services and Management. 2023. Vol. 3. P. 65–73.

16. *Mbonigaba C., Sujatha S., Kumar A.D., Vasuki M.* Leveraging Digital Channels for Customer Engagement and Sales: Evaluating SEO, Content Marketing, and Social Media for Brand Growth // International Journal of Engineering Research and Modern Education. 2024. Vol. 9. P. 32–40.

17. *Lew O.D., Kammerer Y.* Factors influencing viewing behaviour on search engine results pages: A review of eye-tracking research // Behaviour & Information Technology. 2020. Vol. 40. P. 1485–1515.

18. *Wang Q.* Normalization and Differentiation in Google News: A Multi-Method Analysis of the World's Largest News Aggregator: Thesis. Rutgers University, NJ, USA, 2020.

19. *Rahman A.F.R., Alam H., Hartono R.* Content Extraction from HTML Documents // Proceedings of the 1st International Workshop on Web Document Analysis (WDA2001). Seattle, WA, USA, 8 September 2001.

20. *Lima R., Espinasse B., Oliveira H., Pentagrossa L., Freitas F.* Information Extraction from the Web: An Ontology-Based Method Using Inductive Logic Programming // Proceedings of the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence. Herndon, VA, USA, 4–6 November 2013. P. 951–958.

21. *Zheng S., Song R., Wen J.-R.* Template-Independent News Extraction Based on Visual Consistency // Proceedings of the 22nd National Conference on Artificial Intelligence. Vancouver, BC, Canada, 22–26 July 2007. Washington, DC, USA: AAAI Press, 2007. P. 1507–1512.

22. *Zhu W., Dai S., Song Y., Lu Z.* Extracting news content with visual unit of web pages // Proceedings of the 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). Takamatsu, Japan, 1–3 June 2015. P. 1–5.

23. *Gupta S., Kaiser G., Neistadt D., Grimm P.* DOM-based content extraction of HTML documents // Proceedings of the 12th International Conference on World Wide Web. Budapest, Hungary, 20–24 May 2003. P. 207–214.

24. *Mirzaaghaei M., Mesbah A.* DOM-based test adequacy criteria for web applications // Proceedings of the 2014 International Symposium on Software Testing and Analysis. San Jose, CA, USA, 21–26 July 2014. P. 71–81.

25. *Lin J.* Divergence Measures Based on the Shannon Entropy // IEEE Transactions on Information Theory. 1991. Vol. 37, No. 1. P. 145–151.
https://doi.org/10.1109/18.61115

26. *Corander J., Remes U., Koski T.* On the Jensen–Shannon divergence and the variation distance for categorical probability distributions // Kybernetika. 2021. Vol. 57. P. 879–907.

27. *Nielsen F.* Jensen–Shannon divergence and diversity index: Origins and some extensions. Preprint. 2021.

28. *Menéndez M.L., Pardo J.A., Pardo L., Pardo M.C.* The Jensen–Shannon divergence // Journal of the Franklin Institute. 1997. Vol. 334. P. 307–318.

29. Qwen Team. Qwen3-Coder: GitHub repository.
URL: https://github.com/QwenLM/Qwen3-Coder (дата обращения: 11.11.2025).

AUTOMATIC ADDITION OF SEO METADATA TO NEWS ARTICLES USING QWEN-CODER

H. Salem[1] [0000-0002-9143-5231], A. S. Toshchev[2] [0000-0003-4424-6822]

[1]Innopolis University, Innopolis, Russia
[2]Kazan Federal University, Kazan, Russia

[1]h.salem@innopolis.ru, [2]atoschev@kpfu.ru

Abstract

A previously developed pipeline for enriching news articles with structured data is summarized, and an updated configuration is presented in which GPT-3–OpenAI's third-generation natural language processing model – is replaced with Qwen-Coder. As before, the updated enrichment pipeline uses a dataset of 400 pages selected from Google News, a free news aggregator by Google, remains compatible with the Google Rich Results Test (Google's tool for validating eligible structured results), and demonstrates that GPT-3-comparable output quality can be achieved on a low-power desktop PC. We describe how this substitution reduces dependence on paid GPT services and report an evaluation comparing the similarity of outputs produced by Qwen-Coder against the GPT-based baseline. The results also show higher performance of the new algorithm compared with the GPT version. The proposed tools lower the barrier to adopting semantic markup practices and thereby broaden their application in digital journalism. Overall, the findings support Qwen-Coder as a cost-effective alternative to large proprietary models for metadata enrichment tasks.

Keywords: *semantic web, pattern mining, Qwen-Coder, news web pages, readability, structured data.*

REFERENCES

1. *Bashir F., Warraich N.F.* Systematic literature review of Semantic Web for distance learning // Interactive Learning Environments. 2020. Vol. 31. P. 527–543.

2. *Breit A., Waltersdorfer L., Ekaputra F.J., Sabou M., Ekelhart A., Iana A., Paulheim H., Portisch J., Revenko A., Teije A.T., et al.* Combining Machine Learning and Semantic Web: A Systematic Mapping Study // ACM Computing Surveys. 2023. Vol. 55. Art. 313.

3. *Yu L.* Introduction to the Semantic Web and Semantic Web Services. Boca Raton, FL, USA: Chapman and Hall/CRC, 2007.

4. *Sporny M., Longley D., Kellogg G., Lanthaler M., Lindström N.* JSON-LD 1.1: W3C Recommendation. 2020.

5. *Salem H., Salloum H., Orabi O., Sabbagh K., Mazzara M.* Enhancing News Articles: Automatic SEO Linked Data Injection for Semantic Web Integration // Applied Sciences. 2025. Vol. 15. Art. 1262. https://doi.org/10.3390/app15031262

6. OpenAI. GPT-3 powers the next generation of apps. 2021.
URL: https://openai.com/index/gpt-3-apps/

7. *Hui B., Yang J., Cui Z. et al.* Qwen2.5-Coder Technical Report // arXiv. 2024. arXiv:2409.12186. URL: https://arxiv.org/abs/2409.12186

8. *Shadbolt N., Berners-Lee T., Hall W.* The Semantic Web Revisited // IEEE Intelligent Systems. 2006. Vol. 21. P. 96–101.

9. *Poturak M., Keco D., Tutnic E.* Influence of search engine optimization (SEO) on business performance: Case study of private university in Sarajevo // International Journal of Research in Business and Social Science. 2022. Vol. 11. P. 59–68.

10. *Chandrasekaran B., Josephson J.R., Benjamins V.R.* What are ontologies, and why do we need them? // IEEE Intelligent Systems and Applications. 1999. Vol. 14. P. 20–26.

11. *Sporny M., Longley D., Kellogg G., Lanthaler M., Lindström N.* JSON-LD 1.0: W3C Recommendation. 2014.

12. *Adida B., Birbeck M., McCarron S., Pemberton S.* RDFa in XHTML: Syntax and processing: W3C Recommendation. 2008.

13. Rich Results Test. URL: https://search.google.com/test/rich-results

14. *Iqbal M., Khalid M.N., Manzoor A.A., Malik M., Shaikh N.A.* Search Engine Optimization (SEO): A Study of important key factors in achieving a better Search Engine Result Page (SERP) Position // Sukkur IBA Journal of Computing and Mathematical Sciences. 2022. Vol. 6. P. 1–15.

15. *Alfiana F., Khofifah N., Ramadhan T., Septiani N., Wahyuningsih W., Azizah N.N., Ramadhona N.* Apply the Search Engine Optimization (SEO) Method to determine Website Ranking on Search Engines // International Journal of Cyber Services and Management. 2023. Vol. 3. P. 65–73.

16. *Mbonigaba C., Sujatha S., Kumar A.D., Vasuki M.* Leveraging Digital Channels for Customer Engagement and Sales: Evaluating SEO, Content Marketing, and Social Media for Brand Growth // International Journal of Engineering Research and Modern Education. 2024. Vol. 9. P. 32–40.

17. *Lew O.D., Kammerer Y.* Factors influencing viewing behaviour on search engine results pages: A review of eye-tracking research // Behaviour & Information Technology. 2020. Vol. 40. P. 1485–1515.

18. *Wang Q.* Normalization and Differentiation in Google News: A Multi-Method Analysis of the World's Largest News Aggregator: Thesis. Rutgers University, NJ, USA, 2020.

19. *Rahman A.F.R., Alam H., Hartono R.* Content Extraction from HTML Documents // Proceedings of the 1st International Workshop on Web Document Analysis (WDA2001). Seattle, WA, USA, 8 September 2001.

20. *Lima R., Espinasse B., Oliveira H., Pentagrossa L., Freitas F.* Information Extraction from the Web: An Ontology-Based Method Using Inductive Logic Programming // Proceedings of the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence. Herndon, VA, USA, 4–6 November 2013. P. 951–958.

21. *Zheng S., Song R., Wen J.-R.* Template-Independent News Extraction Based on Visual Consistency // Proceedings of the 22nd National Conference on Artificial Intelligence. Vancouver, BC, Canada, 22–26 July 2007. Washington, DC, USA: AAAI Press, 2007. P. 1507–1512.

22. *Zhu W., Dai S., Song Y., Lu Z.* Extracting news content with visual unit of web pages // Proceedings of the 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). Takamatsu, Japan, 1–3 June 2015. P. 1–5.

23. *Gupta S., Kaiser G., Neistadt D., Grimm P.* DOM-based content extraction of HTML documents // Proceedings of the 12th International Conference on World Wide Web. Budapest, Hungary, 20–24 May 2003. P. 207–214.

24. *Mirzaaghaei M., Mesbah A.* DOM-based test adequacy criteria for web applications // Proceedings of the 2014 International Symposium on Software Testing and Analysis. San Jose, CA, USA, 21–26 July 2014. P. 71–81.

25. *Lin J.* Divergence Measures Based on the Shannon Entropy // IEEE Transactions on Information Theory. 1991. Vol. 37, No. 1. P. 145–151.

https://doi.org/10.1109/18.61115

26. *Corander J., Remes U., Koski T.* On the Jensen–Shannon divergence and the variation distance for categorical probability distributions // Kybernetika. 2021. Vol. 57. P. 879–907.

27. *Nielsen F.* Jensen–Shannon divergence and diversity index: Origins and some extensions. Preprint. 2021.

28. *Menéndez M.L., Pardo J.A., Pardo L., Pardo M.C.* The Jensen–Shannon divergence // Journal of the Franklin Institute. 1997. Vol. 334. P. 307–318.

29. Qwen Team. Qwen3-Coder: GitHub repository.
URL: https://github.com/QwenLM/Qwen3-Coder

СВЕДЕНИЯ ОБ АВТОРАХ

САЛЕМ Хамза – аспирант, Университет Иннополис, Лаборатория программной инженерии, г. Иннополис.

Hamza SALEM – PhD student, Innopolis University, Software Engineering Lab, Innopolis.

email: h.salem@innopolis.ru

ORCID: 0000-0002-9143-5231

ТОЩЕВ Александр Сергеевич – доцент, к. н., КФУ, Институт информационных технологий и интеллектуальных систем, Кафедра программной инженерии, г. Казань.

Alexander Sergeevich TOSCHEV – Associate Professor, Ph.D., KFU, Institute of Information Technologies and Intelligent Systems, Department of Software Engineering, Kazan.

email: atoschev@kpfu.ru

ORCID: 0000-0003-4424-6822

Материал поступил в редакцию 18 ноября 2025 года