УДК 004.89
DOI: 10.26102/2310-6018/2026.54.3.008

Метод извлечения информации на основе экстрактивных вопросно-ответных моделей и стратегий оценки и агрегации релевантных фрагментов текста

П.А. Мартынюк✉

Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет), Москва, Российская Федерация

Резюме. В условиях ускоренного роста объемов текстовых данных разнородной структуры особую важность приобретают универсальные подходы к извлечению информации, не зависящие от конкретной структуры и предметной области исходных текстов. Несмотря на широкое распространение больших генеративных языковых моделей, проблема точного и ресурсоэффективного извлечения информации из текстовых данных сохраняет свою актуальность. Генеративные модели, обладая широкими возможностями, зачастую избыточны для решения специализированных задач информационного поиска и могут демонстрировать низкую интерпретируемость получаемых результатов. Настоящее исследование является частью исследовательской работы, направленной на разработку альтернативного метода извлечения информации из неструктурированных текстов с целью формирования структурной модели текстового документа. Предлагаемый подход фокусируется на выделении семантически насыщенных фрагментов текста через анализ релевантности относительно заданных тематических аспектов текста. В рамках данного исследования предлагается метод извлечения информации с использованием экстрактивной вопросно-ответной модели, основанный на многоуровневой агрегации ответов с использованием комбинации стратегий оценки релевантности текстовых фрагментов, семантической кластеризации и выбора результирующего ответа на заданный вопрос. Предлагаемый подход позволяет идентифицировать в тексте слова, наиболее релевантные по отношению к искомым тематическим аспектам, которые впоследствии могут быть использованы для извлечения достоверной информации из документа. В статье представлены результаты эксперимента, подтверждающие эффективность предложенного метода в задаче идентификации семантически релевантных элементов текстового документа. Полученные результаты имеют практическую ценность для разработки систем автоматического построения семантических структур текста и могут быть применены в задачах анализа документов, информационного поиска и интеллектуальной обработки текстовых данных.

Ключевые слова: обработка естественного языка, извлечение информации, неструктурированный текст, вопросно-ответная модель, механизм самовнимания.

Для цитирования: Мартынюк П.А. Метод извлечения информации на основе экстрактивных вопросно-ответных моделей и стратегий оценки и агрегации релевантных фрагментов текста. *Моделирование, оптимизация и информационные технологии.* 2026;14(3). URL: https://moitvivt.ru/ru/journal/article?id=2207 DOI: 10.26102/2310-6018/2026.54.3.008

A method for information extraction based on extractive question-answering models and strategies for evaluating and aggregating relevant text fragments

P.A. Martynyuk✉

Bauman Moscow State Technical University, Moscow, the Russian Federation

Abstract. In the context of accelerated growth of heterogeneous textual data volumes, universal approaches to information extraction that are independent of the specific structure and domain of source texts have become particularly important. Despite the widespread adoption of large generative language models, the problem of accurate and resource-efficient information extraction from textual data remains relevant. While possessing broad capabilities, generative models are often excessive for specialized information retrieval tasks and may demonstrate low interpretability of results. This study is part of research work aimed at developing an alternative method for information extraction from unstructured texts to form a structural model of a text document. The proposed approach focuses on identifying semantically rich text fragments through relevance analysis relative to given thematic aspects of the text. This research presents an information extraction method using an extractive question-answering model, based on multi-level answer aggregation combining strategies for assessing text fragment relevance, semantic clustering, and final answer selection for a given question. The proposed approach enables identification of words in the text that are most relevant to the target thematic aspects, which can subsequently be used to extract reliable information from the document. The article presents experimental results confirming the effectiveness of the proposed method in identifying semantically relevant elements of a text document. The obtained results have practical value for developing automated systems of text semantic structure construction and can be applied in document analysis, information retrieval, and intelligent text processing tasks.

Keywords: natural language processing, information extraction, unstructured text, question-answering model, self-attention mechanism.

For citation: Martynyuk P.A. A method for information extraction based on extractive question-answering models and strategies for evaluating and aggregating relevant text fragments. *Modeling, Optimization and Information Technology*. 2026;14(3). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2207 DOI: 10.26102/2310-6018/2026.54.3.008

Введение

В современном мире в условиях повсеместной цифровизации темпы накопления текстовых данных ускоряются, причем данные обладают разнородной структурой (статьи, отчеты, веб-страницы и пр.). Как следствие, возрастает острая необходимость в универсальных подходах к извлечению информации. Подобные подходы должны обеспечивать обработку текстов независимо от заранее определенной структурной схемы, предметной области или формата представления данных. В частности, это означает, что методы должны быть устойчивы к вариациям жанра, стиля и структуры документа.

Несмотря на бурный рост интереса к большим генеративным языковым моделям (*Large Language Models, LLM*), проблема точного, надежного и ресурсоэффективного извлечения структурированной информации из текста остается критически значимой. Генеративные модели, обладая широкой выразительностью и способностью обрабатывать разнообразные задачи «на лету», могут оказаться избыточными для многих задач информационного поиска [1], где требуется лишь извлечение конкретных элементов (сущностей, отношений, событий). Кроме того, генеративные подходы часто страдают от проблем «галлюцинаций» (ситуации, когда модель предоставляет нерелевантные или неверные факты) [2], а также демонстрируют низкую прозрачность и интерпретируемость получаемых результатов [3]. В этой связи классические и гибридные подходы, которые комбинируют обучение на схемах, модули нейронического распознавания сущностей и отношений, а также нейросетевые структуры с управлением (например, управление с помощью схем, семантические ограничения, промежуточные представления), остаются актуальными. Например, концепция унифицированного извлечения информации (*Universal Information Extraction, IE*) предлагает рассматривать разные задачи *IE* как частные случаи генерации единого структурированного формата из

текста (*text-to-structure*) [4]. Такой подход позволяет объединить извлечение именованных сущностей, отношений, событий и иных компонентов в одну архитектуру, что сокращает накладные расходы на разноструктурные модули и упрощает трансферное обучение между задачами. В задачах извлечения структурированной информации из научных и технических текстов (например, статей, отчетов) особенно актуальна инженерия вывода – обеспечение того, чтобы сгенерированная моделью структура строго удовлетворяла заранее заданной схеме (с принципами валидации). Ярким примером такого подхода является работа [5], где авторы демонстрируют метод совместного извлечения сущностей и отношений из научных текстов и адаптацию больших языковых моделей к этой задаче с учетом специфик научной лексики и структуры статей.

Генеративные методы, формируя выходные данные в свободной текстовой форме, зачастую сталкиваются с проблемами достоверности и верифицируемости извлеченной информации, что особенно критично в прикладных сценариях, требующих точного соответствия исходным данным. В этой связи особый интерес представляют экстрактивные подходы, основанные на выделении фрагментов непосредственно из исходного текста без их переформулирования. Современные экстрактивные методы базируются преимущественно на архитектурах глубоких нейронных сетей, использующих механизмы само-внимания (*self-attention*) и контекстного кодирования, что обеспечивает высокую точность идентификации и локализации релевантных фрагментов текста. Наиболее распространенным направлением является использование моделей типа *BERT* и ее производных (*RoBERTa, DeBERTa, SpanBERT*), способных извлекать сущности и отношения в виде контекстных интервалов без необходимости генерации текста [6]. Подобные модели формируют представления для каждой токен-позиции и обучаются предсказывать границы искомых элементов, что делает их особенно эффективными при решении задач выделения именованных сущностей, отношений и событий. Также в последние годы активно развиваются архитектуры *dual-encoder*, использующие два параллельных кодировщика – для запроса и для корпуса документов. Такие модели обеспечивают компактное представление текста в векторном пространстве, где поиск релевантных фрагментов сводится к задаче нахождения ближайших эмбеддингов [7, 8]. Подобные методы обладают высокой масштабируемостью и позволяют эффективно интегрировать экстрактивное извлечение с системами поиска и вопросно-ответными модулями.

Для повышения устойчивости методов извлечения информации к изменению домена и видов текста в последние годы стали внедряться адаптационные механизмы. Подобные механизмы реализуются как в генеративных, так и в экстрактивных системах, обеспечивая переносимость моделей на новые предметные области и форматы данных без необходимости полного переобучения. Так, например, примером применения адаптационных механизмов в генеративных системах извлечения информации может служить работа [9], в которой авторами предложен подход *ADAPTIVE IE*, реализующий адаптивное извлечение информации «на лету» с использованием генеративных моделей и механизма *human-in-the-loop*. Система динамически формирует и уточняет схемы извлечения на основе пользовательской обратной связи, объединяя автоматическую генерацию вопросов с интерактивной рекластеризацией данных. Предложенный подход обеспечивает гибкость, доменно-независимую адаптацию и повышение точности без необходимости ручной аннотации или фиксированных шаблонов. Примером использования адаптационных компонентов в экстрактивных системах извлечения информации может служить работа [10], в которой представлен метод *DA4JIE* (*Domain Adaptation for Joint Information Extraction*), направленный на решение задачи адаптации к новому домену без использования размеченных данных в рамках совместного

извлечения информации. Предлагаемый подход сочетает два взаимодополняющих модуля: модуль *Instance-relational Domain Adaptation* (*IrDA*), обеспечивающий выравнивание представлений задач между доменами посредством графовых зависимостей, и модуль *Context-invariant Structure Learning* (*CiSL*), формирующий контекстно-независимые структурные представления текста. Разработанная модель демонстрирует способность переносить знания на новые домены без дополнительной аннотации данных.

В настоящем исследовании подход к генеративному извлечению информации не рассматривается ввиду его избыточности для задач извлечения структурированной информации и существенно более высоких требований к вычислительным ресурсам по сравнению с экстрактивными методами, использующими модели-кодировщики [11]. Несмотря на перспективность метода *DA4JIE* и других экстрактивных систем извлечения информации, существующие подходы обладают рядом недостатков. В частности, многие адаптивные экстрактивные модели жестко завязаны на заранее определенные схемы извлечения информации, что ограничивает возможность свободного задания тематических аспектов для построения структурных моделей документов. Целью настоящего исследования является разработка метода извлечения информации на основе использования экстрактивных QA-моделей (*Question Answering*) без необходимости задания жестких схем для извлекаемых компонентов. Требуется, чтобы при использовании подобного метода была возможность свободно задавать тематические аспекты модели для формирования репрезентативных структурных представлений документов, а также, чтобы компоненты метода могли автоматически адаптироваться к предметной области обрабатываемых документов. Разрабатываемый метод рассматривается как ключевой этап при решении более широкой задачи построения структурных моделей текстовых документов, решаемой автором.

Для достижения цели исследования были определены следующие задачи:

1. Формализовать задачу извлечения информации с использованием экстрактивных QA-моделей и последующей агрегацией ответов, полученных из текстовых фрагментов.
2. Предложить стратегии оценки релевантности текстовых фрагментов.
3. Предложить стратегии агрегации ответов QA-моделей для выбора репрезентативного ответа.
4. Подобрать набор данных, содержащий корпус текстовых документов и атрибуты текста для извлечения, и провести экспериментальную проверку предложенного метода с целью выбора наиболее результативных стратегий.
5. Проанализировать практическую применимость предложенного метода.

Научная значимость исследования заключается в разработке метода извлечения информации на основе оценки релевантных фрагментов текста, который расширяет методологию интерпретируемого искусственного интеллекта за счет интеграции обработки скрытых характеристик представлений текстовых данных, поступающих от экстрактивных QA-моделей, с многоуровневой интерпретируемой агрегацией. Предложенный подход служит основой для создания новых принципов структурного представления текстовых документов и открывает перспективы для развития гибридных методов компьютерной лингвистики в условиях роста объемов неструктурированной текстовой информации.

Материалы и методы

Формализация задачи. Пусть задан текст документа D и информационный запрос Q, представленный в виде вопроса на естественном зыке. Текст сегментирован на

фрагменты $S = \{s_1, s_2, ..., s_n\}$ с возможными перекрытиями, где каждый фрагмент s_i представляет собой одно или несколько предложений. Известно, что существует функция QA (соответствует работе экстрактивной QA-модели), которая для каждого фрагмента $s_i \in S$ извлекает ответ $a_i = QA(s_i, Q)$. Требуется среди всех ответов выбрать результирующий ответ a^*, максимально полно удовлетворяющий запросу Q.

Для решения данной задачи необходимо определить следующие функции и стратегии:

1. Функция оценки релевантности ключевых слов:

$$F_k(k_i) \rightarrow \mathbb{R}, \quad (1)$$

где $k_i \in K$ – ключевое слово из множества ключевых слов K, извлеченных из всех фрагментов S;

2. Функция оценки релевантности фрагмента текста:

$$F_s(s_j) \rightarrow \mathbb{R}, \quad (2)$$

где $s_j \in S$ – текстовый фрагмент;

3. Стратегию агрегации ответов и выбора финального ответа:

$$Agg\left(A \middle| F_k(k_i), F_s(s_j)\right) : \{a_1, a_2, \ldots, a_n\} \rightarrow a^*, \quad (3)$$

где с помощью подбора и объединения кандидатов $A = \{a_1, a_2, ..., a_n\}$ формируется финальный ответ a^* с учетом используемых функций $F_k(k_i)$ и $F_s(s_j)$.

Необходимо предложить конкретные функции и стратегии и провести экспериментальное исследование для всех их возможных комбинаций с целью оценки применимости для задачи извлечения информации из неструктурированных текстовых документов.

Схема исследования. Предлагается схема исследования, представленная на Рисунке 1.

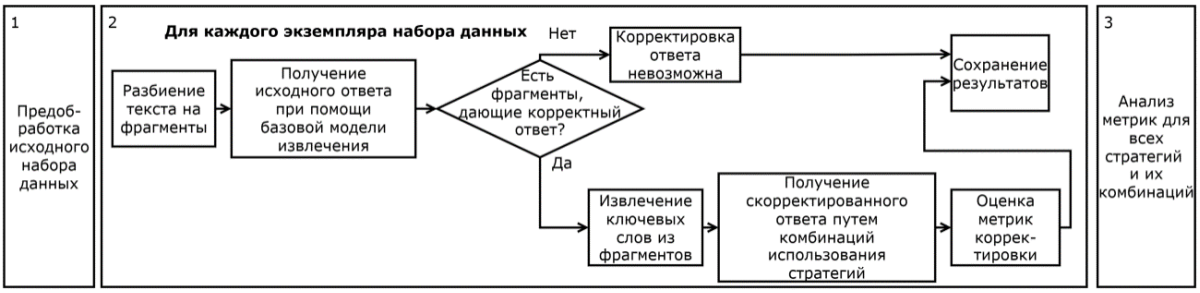


Рисунок 1 – Общая схема исследования
Figure 1 – General scheme of the study

На первом этапе исследования необходимо выполнить предобработку текстовых данных из исходного набора данных. После этого для каждого экземпляра (текста) предлагается выполнить извлечение исходного ответа, причем в первую очередь предлагается использовать некоторый базовый подход к извлечению, а далее попытаться «улучшить» его путем применения предложенных комбинаций стратегий. После этого необходимо оценить эффект от применения стратегий и сделать выводы об их применимости. В качестве базового подхода для извлечения информации предлагается рассмотреть подход, основанный на выделении ответа с наивысшим средним значением семантического сходства, предложенный автором ранее [12]. Также по результатам

анализа предполагается выявить наиболее результативные и практически применимые стратегии из всех предлагаемых.

Набор данных для исследования. Для оценки качества решения задачи извлечения информации из неструктурированных текстов был выбран датасет *SciREX* [13], поскольку он содержит полные тексты научных статей с аннотацией на уровне документа, где каждому фрагменту сопоставлены тематические аспекты – метод, задача, метрика и набор данных. Подобная структура обеспечивает явную связь между текстом и его смысловыми компонентами, что делает SciREX подходящим для выявления тематических связей и дальнейшего построения структурных представлений текста. В ходе предобработки данные были загружены из исходных JSON-файлов, выполнена детокенизация для восстановления исходного текста, а из поля отношений извлечены значения тематических аспектов. Итоговый корпус включает 306 документов, для каждого из которых сформированы текстовые и структурные представления. Для настоящего исследования решено использовать тематический аспект «задача», поскольку он отражает основное направление исследования и позволяет выявить тематическую направленность текста, определяя, к какой области или типу научной проблемы относится документ.

Метрики. Для оценки качества извлечения и последующей корректировки результатов были выбраны две метрики – мера косинусной близости и расстояние Левенштейна. Косинусная мера позволяет оценить семантическое сходство между векторными представлениями исходного и извлеченного фрагментов, что важно при анализе смысловой релевантности. Расстояние Левенштейна, в свою очередь, отражает степень текстового расхождения на уровне символов и служит для выявления лексических и структурных различий. Совместное применение данных метрик обеспечивает комплексную оценку – как смысловой, так и формальной точности извлечения.

Извлечение ключевых слов и оценка их релевантности. В настоящем исследовании предлагается многоэтапная методика извлечения ключевых слов, интегрирующая анализ внутренних параметров механизма внимания нейросетевой вопросно-ответной модели, статистические параметры и контекстные метрики релевантности относительно извлекаемого тематического аспекта.

На первом этапе предлагается идентифицировать семантически значимые токены посредством агрегации весов внимания предобученной модели RoBERTa для задачи QA, где вес значимости $weight_t$ слова t определяется усреднением распределений внимания по всем слоям и «головам» внимания модели-трансформера:

$$weight_k = \frac{1}{L} \cdot \frac{1}{H} \cdot \frac{1}{T} \Sigma_{l=1}^{L} \Sigma_{h=1}^{H} \Sigma_{i=1}^{T} Att[l, h, i, k], \quad (4)$$

где L – количество слоев модели, H – количество голов внимания модели, T – количество токенов контекста, $A[l,h,i,k]$ – количественная мера внимания от токена i к целевому слову k для «головы» h слоя l модели-трансформера.

На втором этапе предлагается выполнить фильтрацию лексических единиц с применением динамического порога для значения *IDF* (*Inverse Document Frequency*), что позволяет исключить термины с низкой дискриминативной способностью. Заключительный этап предполагает вычисление дополнительной метрики тематической релевантности *score_diff*, вычисляемой через сравнительный анализ косинусной близости эмбеддингов извлекаемых слов и эталонными (наименование тематического аспекта) и контрастными (истинный ответ) эмбеддингами:

$$score\_diff_k = cos(E_k, E_{pos}) - cos(E_k, E_{neg}), \quad (5)$$

где E_k – эмбеддинг слова k, E_{pos} – эмбеддинг позитивного эталона (название тематического аспекта), E_{neg} – эмбеддинг негативного эталона (истинный ответ), $cos(a,b)$ – функция косинусного сходства между векторами a и b. Данный подход позволяет идентифицировать слова, которые одновременно репрезентируют содержание исходного текста и при этом соответствуют целевой тематической области.

Оценка релевантности фрагмента текста. В рамках предлагаемой методологии оценки предлагаются три стратегии оценки релевантности текстовых фрагментов. Первая стратегия (*only_score_diff*) основывается исключительно на дискриминативной способности терминов, вычисляемой как разность косинусных сходств. Вторая стратегия (*only_weight*) использует лишь метрику контекстуальной значимости, определяемую через агрегированные веса внимания нейросетевой модели. Третья, гибридная стратегия (*equal_weight_score_diff*) предполагает учет комбинации обеих метрик с коэффициентом балансировки 0,5 и предварительным масштабированием, обеспечивающим равнозначный вклад тематической релевантности и контекстуальной важности в итоговую оценку фрагмента. В общем виде формула для предлагаемых стратегий выглядит следующим образом:

$$F_s(s_j) = \sum_{k \in K_{matched}} [\alpha \cdot weight_k + (1-\alpha) \cdot score\_diff_k] \cdot NormF_{s_j} \cdot DenF_{s_j}, \quad (6)$$

где α – коэффициент балансировки между параметрами ключевых слов, а $NormF$ и $DenF$ – фактор нормализации и фактор плотности соответственно. Для первой стратегии $\alpha = 1$, для второй $\alpha = 0$, а для гибридной стратегии $\alpha = 0{,}5$ соответственно.

В формуле (6) фактор нормализации по длине фрагмента $NormF$ вводится для того, чтобы избавиться от систематического смещения в пользу протяженных текстовых фрагментов. Без данной корректировки возникает статистический артефакт, при котором более длинные фрагменты получают необоснованное преимущество исключительно благодаря большей вероятности содержания ключевых терминов в силу своего объема, что не имеет необходимой корреляции с действительной релевантностью содержания:

$$NormF = 1 + \frac{|K_{matched}|}{0{,}1 \cdot T_{s_j}}, \quad (7)$$

где $K_{matched} \subseteq K$ – множество найденных ключевых терминов во фрагменте, T_{sj} – количество токенов во фрагменте s_j. Множитель 0,1 в факторе нормализации установлен эмпирически как оптимальное значение для балансировки влияния длины фрагмента: большие значения приводят к избыточному усилению коротких текстов, меньшие – нивелируют преимущества компактных релевантных сегментов.

В свою очередь, фактор плотности ключевых терминов $DenF$ предназначен для усиления значимости текстовых сегментов с высокой концентрацией релевантной лексики. Психолингвистические исследования подтверждают, что высокая плотность тематически значимых единиц является индикатором смысловой целостности и сфокусированности содержания на целевой проблематике [14], что позволяет количественно оценивать насыщенность текста содержательными лексемами, связанными с его тематической областью. Ограничение величины данного коэффициента верхней границей 2,0 предотвращает гиперкомпенсацию для чрезмерно кратких фрагментов со случайными лексическими совпадениями, обеспечивая устойчивость метрики к шумовым воздействиям:

$$DenF = min\left(2{,}0; 1 + \frac{|K_{matched}|}{0{,}2 \cdot W_f}\right), \quad (8)$$

где W_f – количество слов во фрагменте s_j. Множитель 0,2 в факторе плотности отражает пороговое значение концентрации ключевых терминов (20 % от общего объема текста),

при котором фрагмент считается тематически сфокусированным. Значение выбрано на основании лингвостатистических наблюдений распределения значимой лексики: в эмпирическом исследовании [15] частота слов, оцениваемых экспертами как «высоко характерные» для тематической области, составляла порядка 50 %, 30 % и 20 % для различных тем, при этом минимальное значение – около 20 % – может рассматриваться как нижний порог тематической фокусировки.

Совместное применение данных корректирующих коэффициентов реализует принцип балансировки между репрезентативностью и концентрацией семантического содержания, где оптимальными признаются фрагменты, сочетающие достаточный объем для контекстуальной полноты с высокой плотностью релевантной информации.

Стратегия агрегации ответа. Предлагается агрегация множественных ответов через их семантическую кластеризацию. Исходная гипотеза предполагает, что релевантные ответы образуют компактные семантические кластеры в пространстве эмбеддингов, тогда как ошибочные и шумовые ответы распределяются случайным образом. Кластеризация осуществляется методом одиночной связи с пороговым критерием семантического сходства, что позволяет выявлять устойчивые семантические паттерны среди вариантов ответов. В результате формируется множество кластеров $C = \{C_1, C_2, ..., C_M\}$, где $C_i = \{a \in A \mid sim(a, centroid_i) \geq \theta\}$, θ – порог семантического сходства (для эксперимента используется значение $\theta = 0{,}75$, значение подобрано экспертно), $centroid_i$ – центроид кластера C_i, $sim(a,b)$ – функция косинусного сходства эмбеддингов.

После выполнения кластеризации необходимо выявить наиболее релевантный кластер для поиска результирующего ответа. Предлагается три стратегии выбора такого кластера. Стратегия наивысшего среднего качества (*highest_avg_score*) предполагает выбор кластера с максимальным усредненным показателем релевантности фрагментов-источников:

$$C_{sel} = argmax_{C_i \in C} \left[\left(\frac{1}{|C_i|} \right) \sum_{\substack{a \in C_i \\ a = QA(s)}} F_s(s) \right]. \quad (9)$$

Данный подход оптимизирует объективное качество содержания, но может игнорировать кластеры с экстремально релевантными единичными ответами.

Стратегия взвешенной оценки (*weighted_score*) комбинирует размер кластера и среднее качество ответов, обеспечивая баланс между консенсусной поддержкой и содержательной ценностью. Мультипликативная функция гарантирует, что предпочтение отдается кластерам, одновременно демонстрирующим как количественную представительность, так и высокое семантическое качество:

$$C_{sel} = argmax_{C_i \in C} \left[|C_i| \cdot \left(\frac{1}{|C_i|} \right) \sum_{\substack{a \in C_i \\ a = QA(s)}} F_s(s) \right] = argmax_{C_i \in C} \left[\sum_{\substack{a \in C_i \\ a = QA(s)}} F_s(s) \right]. \quad (10)$$

Стратегия семантической когерентности (*highest_cohesion*) выделяет кластер с максимальным внутренним сходством эмбеддингов ответов. Данный подход основывается на предположении, что высокая семантическая сплоченность коррелирует с концептуальной целостностью и точностью формулировок:

$$C_{sel} = argmax_{C_i \in C} \left[\left(\frac{2}{|C_i|(|C_i|-1)} \right) \sum_{i=1}^{|C_i|} \sum_{j=i+1}^{|C_i|} sim(a_i, a_j) \right]. \quad (11)$$

Далее из выбранного кластера необходимо выделить результирующий ответ a^*, который наилучшим образом обобщал бы все ответы кластера. В настоящем исследовании предлагаются и рассматриваются следующие стратегии отбора ответа.

Стратегия максимальной релевантности источника (*highest_chunk_score*) выбирает ответ $a^* \in C_{sel}$, полученный из фрагмента s, $a^* = QA(s)$, с наивысшей оценкой $F_s(s)$, она позволяет приоритизировать содержательную точность над формальными характеристиками. Стратегия семантической репрезентативности (*highest_sim*) отдает предпочтение ответу с наибольшим средним сходством с другими элементами кластера. Данный подход максимизирует консенсусный потенциал ответа, обеспечивая его максимальную репрезентативность для всей семантической группы. Комбинированная стратегия (*combined*) объединяет показатели релевантности источника и семантической репрезентативности, что позволяет одновременно учитывать оба показателя. В общем виде стратегии для выбора ответа можно формализовать следующим образом:

$$a^* = argmax_{a \in C_{sel}} \left[\beta \, F_s(s)_{a=QA(s)} + (1-\beta) \left(\frac{1}{|C_{sel}|-1} \right) \sum_{\substack{b \neq a \\ a,b \in C_{sel}}} sim(a,b) \right], \quad (12)$$

где β – коэффициент балансировки между оценкой фрагмента и средним значением сходства ответов в выбранном кластере. Для стратегии *highest_chunk_score* $\beta = 1$, для стратегии *highest_sim* $\beta = 0$ и для стратегии *combined* $\beta = 0{,}5$.

В рамках настоящего исследования предлагается проведение комплексного экспериментального исследования, направленного на верификацию эффективности предложенной многоуровневой методологии агрегации ответов. Эксперимент предусматривает сравнительный анализ всех 27 стратегий (3×3×3) на репрезентативной выборке текстовых данных.

Результаты

Экспериментальное исследование выявило потенциал улучшения качества ответов у 281 из 306 примеров (91,83 %). Анализ применимости предлагаемого метода проводился поэтапно: оценивалось влияние отдельных стратегий на каждом уровне обработки, а затем проанализированы их комбинации. В Таблицах 1–3 представлены сравнительные результаты стратегий оценки фрагментов, выбора кластера и финального ответа, где наилучшие показатели выделены полужирным шрифтом.

Таблица 1 – Сравнение стратегий оценки фрагмента
Table 1 – Comparison of chunk evaluation strategies

Стратегия оценки фрагмента	**Косинусная мера**		**Расстояние Левенштейна**		**Обе метрики**	
	% улуч-шений	% ухуд-шений	% улуч-шений	% ухуд-шений	% улуч-шений	% ухуд-шений
only_score_diff	50,42	28,35	42,31	33,21	35,51	19,77
only_weight	**51,09**	**27,80**	**43,10**	**32,78**	**36,30**	**19,45**
equal_weight_score_diff	50,42	28,31	42,31	33,18	35,51	19,73

На уровне вычисления весовых коэффициентов фрагментов наилучшие результаты продемонстрировала стратегия *only_weight*, основанная исключительно на контекстуальной значимости терминов, которая определяется на основании значений в слоях механизма внимания модели-трансформера. На этапе выбора результирующего кластера ответов лучшей оказалась стратегия *weighted_score*, интегрирующая количественные (размер) и качественные (оценка фрагментов) показатели кластеров.

Таблица 2 – Сравнение стратегий выбора результирующего кластера
Table 2 – Comparison of strategies for selecting the resulting cluster

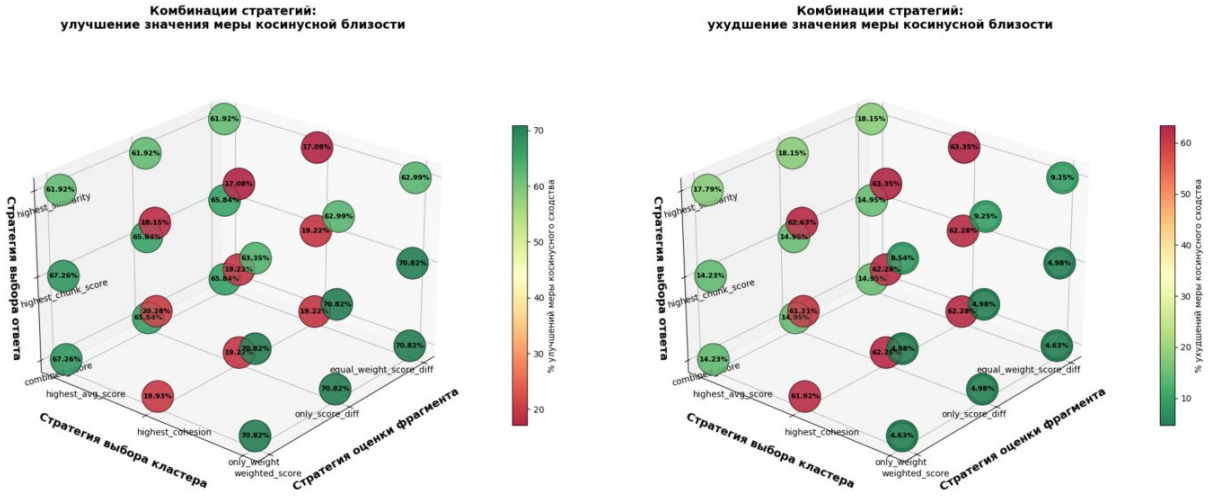
Стратегия выбора кластера	Косинусная мера		Расстояние Левенштейна		Обе метрики	
	% улучшений	% ухудшений	% улучшений	% ухудшений	% улучшений	% ухудшений
highest_avg_score	64,85	15,82	48,83	29,10	45,75	12,38
highest_cohesion	18,82	62,4	27,52	49,82	11,23	41,87
weighted_score	**68,25**	**6,25**	**51,36**	**20,25**	**50,34**	**4,71**

Таблица 3 – Сравнение стратегий выбора результирующего ответа
Table 3 – Comparison of strategies for selecting the resulting answer

Стратегия выбора ответа	Косинусная мера		Расстояние Левенштейна		Обе метрики	
	% улучшений	% ухудшений	% улучшений	% ухудшений	% улучшений	% ухудшений
highest_chunk_score	**52,23**	**27,20**	43,89	32,31	37,41	19,02
highest_similarity	47,49	30,05	39,66	34,84	32,38	21,04
combined_score	52,19	**27,20**	**44,17**	**32,03**	**37,52**	**18,90**

На заключительном этапе выбора финального ответа внутри кластера наилучшие результаты показала стратегия *combined_score*, объединяющая показатели релевантности источника (оценка фрагмента) и семантической репрезентативности ответа (семантическая близость с другими ответами из кластера).

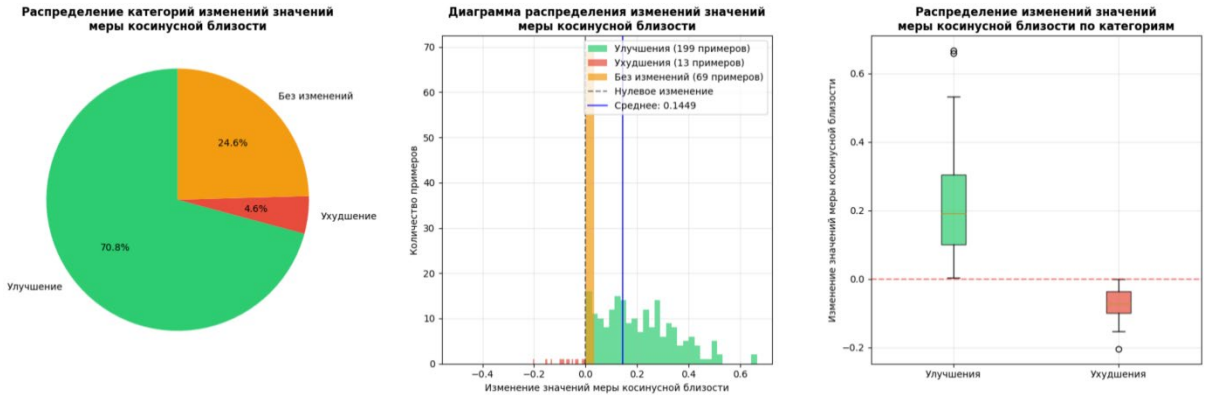
Дополнительно проведено сравнительное исследование комбинаций стратегий с оценкой их влияния на улучшение и ухудшение качества ответов. Сравнение эффективности стратегий выполнено исключительно по метрике косинусной близости, поскольку она отражает семантическое сходство между эталонными и полученными ответами на уровне смыслового содержания. Результаты исследования для метрики косинусной близости представлены на Рисунке 2.



а) *б*)

Рисунок 2 – Сравнение результативности комбинаций стратегий по изменению значения меры косинусной близости: *а* – улучшение; *б* – ухудшение
Figure 2 – Comparison of the effectiveness of combinations of strategies for changing the value of the cosine similarity measure: *a* – improvement; *b* – deterioration

Анализ выявил, что наилучшими комбинациями оказались *only_weight* + *weighted_score* + *combined_score* и *equal_weight_score_diff* + *weighted_score* + *combined_score*. Обе комбинации обеспечили значительное улучшение качества ответов в 70,82 % случаев от общего числа примеров с потенциалом оптимизации и при этом ухудшили качество в минимуме примеров (4,63 %). Оптимальной была признана комбинация *only_weight* + *weighted_score* + *combined_score*, поскольку она демонстрирует более высокое среднее изменение значения косинусной близости – 0,1449 против 0,1441 у стратегии *equal_weight_score_diff* + *weighted_score* + *combined_score*. Дополнительно выполнен анализ результатов применения наилучшей стратегии, на Рисунке 3 представлены графики распределения изменений значений косинусной близости для данной комбинации стратегий, наглядно демонстрирующие, что в подавляющем большинстве случаев наблюдается существенный положительный сдвиг метрики.



а) *б*) *в*)

Рисунок 3 – Графики распределения изменений значений косинусной близости для оптимальной комбинации стратегий: *а* – по категориям; *б* – диаграмма распределения; *в* – диаграмма размаха значений

Figure 3 – Distribution graphs of changes in cosine similarity values for the optimal combination of strategies: *a* – by category; *b* – distribution diagram; *c* – boxplot diagram

Обсуждение

Преимущество стратегии оценки фрагмента текста *only_weight* свидетельствует о том, что семантические паттерны, выявляемые механизмом внимания трансформера в исходном контексте, обеспечивают более надежную базу для оценки релевантности фрагментов, чем метрики, основанные на внешнем семантическом сравнении. Данный результат указывает на перспективность дальнейшего исследования возможностей прямого использования внутренних представлений моделей-трансформеров для задач оценки качества текстового фрагмента. Стратегия выбора результирующего кластера *weighted_score* продемонстрировала наивысшую результативность при выборе кластера благодаря своей способности одновременно учитывать количественный показатель консенсусной поддержки (размер кластера) и качественную характеристику содержательной релевантности (совокупную оценку фрагментов-источников). Данный подход обеспечивает оптимальный баланс между репрезентативностью и качеством, исключая риски выбора чрезмерно гетерогенных кластеров большого размера или малоустойчивых кластеров с экстремальными точечными оценками. Эффективность стратегии выбора результирующего ответа *combined_score* на этапе финального выбора ответа объясняется синергетическим взаимодействием объединяемых метрик. Анализ

результатов показывает, что показатель релевантности источника гарантирует содержательную достоверность ответа, в то время как метрика семантической репрезентативности обеспечивает его максимальное соответствие тематическому профилю кластера. Экспериментальные данные подтверждают, что такой комплексный подход позволяет стабильно идентифицировать ответы, которые одновременно характеризуются высокой точностью и оптимально отражают смысловое ядро кластера. Сравнительный анализ всех возможных комбинаций стратегий выявил четкую градацию их эффективности, что позволило обоснованно выбрать оптимальную конфигурацию, которая продемонстрировала превосходство по ключевым метрикам качества: максимальный процент улучшения ответов (70,82 %), минимальный процент ухудшения (4,63 %) и наибольшее среднее увеличение значения косинусной близости (0,1449).

Заключение

Исследование подтвердило практическую применимость предложенного метода извлечения информации на основе использования стратегий оценки релевантности текстовых фрагментов, семантической кластеризации ответов, их последующего ранжирования и выбора результирующего ответа. Экспериментально подтверждено, что оптимальная комбинация стратегий *only_weight* + *weighted_score* + *combined_score* обеспечивает улучшение качества ответов с минимальным процентом ухудшения. В качестве дальнейшего направления исследований планируется разработка механизма динамического накопления и аккумуляции ключевых слов в процессе работы системы, позволяющего формировать и постоянно обновлять корпус релевантных терминов для последующего использования при извлечении информации из разнородных документальных массивов. Такой подход позволит создать систему, адаптирующуюся к тематическим особенностям обрабатываемых документов и повышающую точность извлечения релевантной информации за счет накопления семантически верифицированных ключевых элементов.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Xu D., Chen W., Peng W., et al. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*. 2024;18(6). https://doi.org/10.1007/s11704-024-40555-y
2. Huang L., Yu W., Ma W., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. 2025;43(2). https://doi.org/10.1145/3703155
3. Zhao H., Chen H., Yang F., et al. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology.* 2024;15(2). https://doi.org/10.1145/3639372
4. Cong X., Yu B., Fang M., et al. Universal information extraction with meta-pretrained self-retrieval. In: *Findings of the Association for Computational Linguistics: ACL 2023, 09–14 July 2023, Toronto, Canada*. Association for Computational Linguistics; 2023. P. 4084–4100. https://doi.org/10.18653/v1/2023.findings-acl.251
5. Dagdelen J., Dunn A., Lee S., et al. Structured information extraction from scientific text with large language models. *Nature Communications*. 2024;15. https://doi.org/10.1038/s41467-024-45563-x
6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, NAACL-HLT 2019: Volume 1, 02–07 June 2019, Minneapolis, MN, USA. Association for Computational Linguistics; 2019. P. 4171–4186.

7. Karpukhin V., Oguz B., Min S., et al. Dense Passage Retrieval for Open-Domain Question Answering. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, 16–20 November 2020, Online.* Association for Computational Linguistics; 2020. P. 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550
8. Izacard G., Grave E. *Distilling Knowledge from Reader to Retriever for Question Answering.* arXiv. URL: https://doi.org/10.48550/arXiv.2012.04584 [Accessed 12th January 2026].
9. Mondal I., Yuan M., Natarajan A., et al. ADAPTIVE IE: Investigating the Complementarity of Human-AI Collaboration to Adaptively Extract Information on-the-fly. In: *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, 19–24 January 2025, Abu Dhabi, UAE.* Association for Computational Linguistics; 2025. P. 5870–5889.
10. Ngo N.T., Min B., Nguyen Th.H. Unsupervised domain adaptation for joint information extraction. In: *Findings of the Association for Computational Linguistics: EMNLP 2022, 07–11 December 2022, Abu Dhabi, UAE*. Association for Computational Linguistics; 2022. P. 5894–5905. https://doi.org/10.18653/v1/2022.findings-emnlp.434
11. Arzideh K., Schäfer H., Allende-Cid H., et al. From BERT to generative AI – Comparing encoder-only vs. large language models in a cohort of lung cancer patients for named entity recognition in unstructured medical reports. *Computers in Biology and Medicine*. 2025;195. https://doi.org/10.1016/j.compbiomed.2025.110665
12. Березкин Д.В., Козлов И.А., Мартынюк П.А., Панфилкин А.М. Метод создания структурных моделей текстовых документов с использованием нейронных сетей. *Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика.* 2023;12(1):28–45. (На англ.). https://doi.org/10.14529/cmse230102
Berezkin D.V., Kozlov I.A., Martynyuk P.A., Panfilkin A.M. A method for creating structural models of text documents using neural networks. *Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering.* 2023;12(1):28–45. https://doi.org/10.14529/cmse230102
13. Jain S., Van Zuylen M., Hajishirzi H., Beltagy I. SciREX: A challenge dataset for document-level information extraction. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, 05–10 July 2020, Online.* Association for Computational Linguistics; 2020. P. 7506–7516. https://doi.org/10.18653/v1/2020.acl-main.670
14. Graesser A.C., McNamara D.S., Louwerse M.M., Cai Zh. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers.* 2004;36(2):193–202. https://doi.org/10.3758/BF03195564
15. Prentice Sh., Knight J., Rayson P., Haj M.E., Rutherford N. Problematising characteristicness: a biomedical association case study. *International Journal of Corpus Linguistics.* 2021;26(3):305–335. https://doi.org/10.1075/ijcl.19019.pre

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Мартынюк Полина Антоновна, ассистент кафедры «Компьютерные системы и сети», Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет), Москва, Российская Федерация.
e-mail: martynyuk.pa@bmstu.ru
ORCID: 0000-0002-2429-1805

Polina A. Martynyuk, Assistant at the Department «Computer Systems and Networks», Bauman Moscow State Technical University, Moscow, the Russian Federation.

Статья поступила в редакцию 30.01.2026; одобрена после рецензирования 07.03.2026; принята к публикации 17.03.2026.

The article was submitted 30.01.2026; approved after reviewing 07.03.2026; accepted for publication 17.03.2026.