ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

УДК 81'322.2

Василевская Валерия Михайловна
аспирант кафедры теоретической
и прикладной лингвистики
Белорусский государственный
университет иностранных языков
г. Минск, Беларусь

Valeria Vasilevskaya
PhD Student of the Department
of Theoretical and Applied Linguistics
Belarusian State University
of Foreign Languages
Minsk, Belarus
valeria-vas90@mail.ru

ОСОБЕННОСТИ ПЕРЕДАЧИ СТРУКТУРЫ РУССКОЯЗЫЧНОЙ НАУЧНОЙ СТАТЬИ ПРИ РЕФЕРИРОВАНИИ БОЛЬШИМИ ЯЗЫКОВЫМИ МОДЕЛЯМИ
(на материале статей разного объема
из области информационно-коммуникационных технологий)

STRUCTURAL REPRESENTATION IN LLM-GENERATED SUMMARIES OF RUSSIAN ACADEMIC TEXTS
(Based on Articles of Various Lengths in the Field
of Information and Communication Technologies)

Настоящая работа посвящена исследованию структурных особенностей автоматических рефератов научных статей в области информационно-коммуникационных технологий, сгенерированных большими языковыми моделями. Проведен сопоставительный анализ структуры полученных рефератов с оригинальными текстами статей с фокусом на полноту передачи микротем, значимых для понимания содержания. На основе полученных данных выдвинуты гипотезы о возможных факторах, влияющих на способность языковых моделей распознавать структурные компоненты текста и формировать структуру реферата.

Ключевые слова: *реферирование*; *структура*; *большие языковые модели*; *научный дискурс*; *информационно-коммуникационные технологии*.

This paper is devoted to the study of structural features of automatic abstracts of scientific articles in the field of information and communication technologies generated by large language models. A comparative analysis was conducted between the generated summaries and the original research articles, focusing on the completeness of microtopic representation essential for content comprehension. Based on the data obtained, hypotheses were put forward about possible factors influencing the ability of language models to recognize structural components of the text and form the structure of the abstract.

Key words: *summarization*; *structure*; *large language models*; *scientific discourse*; *information and communication technologies*.

В условиях современной научной коммуникации нейросетевые технологии становятся популярным средством автоматического реферирования научных текстов. Особенно актуальны эти технологии в условиях экспонен-

циального роста объема научных публикаций, когда традиционные методы ручного реферирования оказываются недостаточно оперативными.

Несмотря на значительные достижения в задачах генерации, классификации и предсказания, большие языковые модели по-прежнему характеризуются как «черные ящики» [1; 2, с. 146] ввиду отсутствия прозрачности в механизмах принятия решений. Как отмечает А. Г. Кузнецов, «нейросети непрозрачны или неинтерпретируемы не только для исследователей или аудиторов, но и для самих их создателей» [3, с. 174]

При работе с узкоспециализированными научными текстами, в частности из области информационно-коммуникационных технологий (ИКТ), языковые модели сталкиваются с рядом лингвистических трудностей, связанных как со спецификой русского языка в целом [4; 5], так и с особенностями соответствующего научного дискурса, включая: 1) передачу сложной терминологии и распознавание контекстно-зависимых значений терминов, которые могут варьироваться в разных предметных областях; 2) интерпретацию специфических аббревиатур при недостаточном контексте; 3) представление формализованных элементов (формул, таблиц, алгоритмов); 4) восстановление сложных концептуальных и логико-семантических связей [5].

Отдельную проблему составляет сохранение строгой логико-композиционной структуры научного текста, которая служит основополагающим признаком академического дискурса [6].

Цель настоящего исследования заключается в выявлении особенностей автоматического реферирования русскоязычных научных текстов ИКТ-тематики нейросетевыми моделями. Основное внимание уделяется структуре вторичных текстов, а также факторам, определяющим стратегию генерации, реализуемую языковыми моделями.

В качестве инструментария были выбраны три большие языковые модели (Large Language Model, LLM): DeepSeek-V3, GPT-4, Qwen2.5-Max. Критериями выбора являлись: 1) доступность в белорусском сегменте интернета; 2) бесплатный доступ к функционалу; 3) поддержка загрузки PDF-файлов для обработки.

Эмпирическую базу составили три статьи, отобранные из научной электронной библиотеки «КиберЛенинка» [7], посвященные применению искусственного интеллекта в обработке естественного языка. Для каждой статьи были сгенерированы по три автоматических реферата – по одному от каждой модели, что позволило в сумме получить девять рефератов.

Критерии отбора статей включали:

1) объем и степень структурированности (статья небольшого объема 13 437 знаков с пробелами, подробно структурированная; средняя – 34 016 знаков с минимальной структурой; крупная – 45 341 знак с детализированной иерархией разделов);

2) наличие элементов, типичных для научных текстов в ИКТ-дискурсе: формул, таблиц, алгоритмов, графиков, насыщенности аббревиатурами преимущественно на латинице.

Принципиальным условием исследования стало моделирование типичной пользовательской ситуации: отсутствие предварительной обработки исходного текста, постобработки сгенерированного реферата и использование минимально детализированного запроса к модели (без составления сложных промптов).

Структура исходных статей соответствовала схеме: метаданные (заголовок, аннотация, ключевые слова), введение, основная часть, заключение, список литературы. Основная часть варьировалась по структуре и включала различное количество разделов и подразделов.

В ходе исследования каждой из выбранных LLMs поочередно предлагались три научные статьи с запросом на автоматическое реферирование при условии сохранения всей ключевой информации, в том числе данные из таблиц и графиков. Полученные рефераты были разбиты на микротемы, которые затем сопоставлялись с микротемами, дифференцированными в оригинальных текстах.

1. Анализ полученных рефератов статьи небольшого объема, содержащей подробную структурацию на подразделы

Структура оригинального текста включала следующие выделенные автором компоненты: 1. Метаданные (информация об авторах, название статьи, аннотация, ключевые слова); 2. Введение (три микротемы); 3. Основная часть (два раздела, содержащие соответственно четыре и две микротемы); 4. Заключение (три микротемы); 5. Список литературы.

Результаты идентификации метаданных:

Наивысшую точность продемонстрировала DeepSeek-V3, корректно передав всю необходимую информацию. GPT-4 экстрактивно включила в реферат название статьи, аннотацию и ключевые слова, но полностью элиминировала данные об авторах. Qwen2.5-Max, напротив, отобразила информацию об авторах и частично содержание аннотации, но упустила название статьи.

Структура введения включала три микротемы: 1) определение предмета исследования, 2) представление мировой статистики по теме, 3) обоснование актуальности и характеристика современного состояния вопроса.

Все модели корректно определили предмет исследования и распознали наличие статистических данных. Однако DeepSeek-V3 и GPT-4 испытывали затруднения при извлечении информации об актуальности, что, вероятно, связано с ее неявной формулировкой: слово *актуальность* в тексте отсутствовало, и это могло затруднить идентификацию значимости темы без явных лингвистических маркеров.

Реферирование основной части продемонстрировало стремление моделей, особенно DeepSeek-V3 и GPT-4, к сохранению авторской логики названий разделов.

Все четыре микротемы раздела 1 были корректно отражены только в реферате, сгенерированном GPT-4. DeepSeek-V3 выделила лишь две микротемы, при этом одна из них была ошибочно фрагментирована и дополнена галлюцинациями (некорректными, несуществующими элементами). Резуль-

таты Qwen2.5-Max по этому разделу оказались неудовлетворительными: модель не смогла адекватно структурировать насыщенный микротемами фрагмент текста при отсутствии явных авторских акцентов.

Раздел 2 включал две микротемы: первая представляла собой маркированный список с авторским шрифтовым выделением, вторая содержала формулу. Все LLMs корректно идентифицировали обе микротемы, включая структурную специфику оформления научного текста, содержащего средства наглядности.

Заключение состояло из трех микротем, которые все модели отразили в целом корректно.

Список литературы был номинирован в рефератах, сгенерированных DeepSeek-V3 и GPT-4 с указанием общего количества источников.

2. Анализ полученных рефератов статьи среднего объема, имеющей небольшое число структурных частей

Структура оригинальной статьи включала следующие элементы: 1. Метаданные (ФИО авторов, название статьи, аннотация, ключевые слова); 2. Введение (четыре микротемы); 3. Основная часть (два раздела, содержащие соответственно по три и четыре микротемы); 4. Заключение (одна микротема); 5. Библиографический список; 6. Повторные метаданные (расширенная информация об авторах).

Обработка метаданных показала, что ни одна из моделей не распознала авторов статьи. Это, вероятно, обусловлено нестандартным расположением данной информации: фамилии и инициалы авторов даны до названия статьи, а их полные сведения – после библиографического списка.

Дополнительная структурная особенность рассматриваемой статьи – чрезмерно объемная и детализированная аннотация, которая по сути представляет собой краткое изложение содержания всей работы. В результате при ее обработке LLMs интерпретировали ряд ключевых фрагментов основного текста как уже представленные ранее. В частности, все модели полностью элиминировали микротемы введения, в том числе подробную информацию о цели и задачах исследования.

Основная часть включала два раздела, при этом первый раздел значительно превышал второй по объему.

Раздел 1 содержал три микротемы. Первые две насыщены англоязычными аббревиатурами (*ANN, RNN, CNN*), различающимися формально лишь одной буквой и не сопровождающимися расшифровкой. Это вызвало затруднения у моделей. В частности, в реферате, сгенерированном DeepSeek-V3, была полностью опущена меньшая по объему микротема, посвященная ANN и RNN, в пользу более подробно представленной микротемы о CNN. В рефератах от GPT-4 и Qwen2.5-Max данные аббревиатуры упоминаются лишь кратко и обобщенно. Третья микротема этого раздела включала иллюстрацию, формулу и список, выполняющие функцию визуальных и структурных ориентиров. Благодаря этому все модели корректно распознали данный фрагмент текста.

Раздел 2 содержал четыре микротемы, но при этом имел небольшой объем. Как и в случае с первой проанализированной статьей, отсутствие явных авторских акцентов (например, шрифтового выделения) в насыщенном микротемами сплошном тексте затруднило моделям выделение иерархии смысловых единиц. Ни одна LLM не справилась с задачей полной и точной передачи содержательных единиц раздела 2.

Заключение представляло собой обобщение основных положений статьи и частично дублировало формулировки из аннотации. Несмотря на это, все модели корректно идентифицировали заключение как значимый фрагмент и включили его в рефераты.

Библиографический список был полностью исключен из всех рефератов. Примечательно, что в тех статьях, где данный раздел обозначался как *Список литературы*, подобной проблемы не возникло. Это позволяет предположить, что слово *литература* в наименовании раздела выступает для моделей в качестве маркера, помогающего идентифицировать данный раздел как значимую часть научного текста, подлежащую отражению в реферате.

3. Анализ полученных рефератов статьи большого объема с детально представленной структурой

Структура оригинальной статьи включала следующие части: 1. Метаданные (информация об авторе, название статьи, аннотация, ключевые слова); 2. Введение (две микротемы); 3. Основная часть (три раздела, два из которых имеют несколько подразделов); 4. Заключение (одна микротема); 5. Выражение благодарности; 6. Список литературы.

Статья отличалась высокой содержательной насыщенностью каждого подраздела, а также значительным количеством аббревиатур, числовых данных, выделенных шрифтом примеров.

Необходимо отметить, что для текстов большого объема с изначально развитой иерархической структурой характерна высокая сохранность структурных элементов в генерируемых рефератах. В частности, GPT-4 продемонстрировала абсолютно идентичное авторскому деление на разделы и подразделы.

Анализ показал, что все LLMs верно передали метаданные в рефераты. Наиболее полно с этой задачей справилась GPT-4, сохранив в сгенерированном тексте даже ключевые слова, обычно не входящие в состав автоматически создаваемых нейросетевыми моделями рефератов.

В ходе дальнейшей обработки статьи DeepSeek-V3 и Qwen2.5-Max объединили краткое введение и поясняющий предмет исследования раздел 1 основной части, утеряв при этом одну из важных микротем введения.

В разделе 2 подраздел 2.1 «Проведение международных конференций-соревнований» посвящен истории поиска решения поставленной задачи и содержит важные пояснения. Проанализировав название подраздела, DeepSeek-V3 и Qwen2.5-Max классифицировали его как несущественный для

понимания сути исследования и практически полностью элиминировали при реферировании, перейдя непосредственно к следующей за ним части о методологии исследования. Далее обе вышеупомянутые LLMs объединили небольшие подразделы 2.2 и 2.4 с более крупным подразделом 2.3, изменив авторский порядок следования микротем, однако без нарушения общей логики изложения. DeepSeek-V3 исключила из реферата подраздел 2.5 «Основные задачи, связанные с ТЕ», вероятно, идентифицировав его как дублирование подраздела 2.3 «Задача извлечения и нормализации ТЕ».

Раздел 3 (однотипные подразделы 3.1–3.5) посвящен системам и методам решения обозначенной в статье задачи. Рассмотрению каждой системы либо метода отведен отдельный подраздел (всего пять подразделов), название которого совпадает с наименованием описываемого в нем феномена. При обработке данного раздела все использованные модели столкнулись со сложностями. Так, DeepSeek-V3 включила в реферат только четыре из пяти позиций; GPT-4 при полном сохранении авторской структуры номинировала все пять позиций с краткими пояснениями, недостаточными для понимания сути; Qwen2.5-Max в произвольном порядке номинировала три позиции без каких-либо пояснений.

Задача идентификации и реферирования заключения была решена всеми LLMs успешно.

Структурной особенностью являлось наличие в конце статьи авторской благодарности, сопровождавшейся шрифтовым акцентом. В отличие от проигнорировавших данную часть DeepSeek-V3 и Qwen2.5-Max, GPT-4 классифицировала ее как значимую и включила в реферат экстрактивным способом.

Структурный анализ автоматически сгенерированных LLM-рефератов русскоязычных научных текстов ИКТ-тематики позволяет выдвинуть ряд **гипотез** о наличии прямой корреляции между степенью структурной организованности исходного текста (в частности – четкостью выделения автором ключевых структурно-смысловых элементов) и качеством результирующего реферата. Ниже представлены основные положения, выявленные в результате анализа.

1. Логически и стилистически четкие авторские акценты повышают качество автоматического реферирования.

Если в тексте научной статьи соблюдены дискурсивные нормы научного стиля и явно обозначены смысловые опорные точки, LLMs демонстрируют более точную обработку информации и сниженную вероятность генерации так называемых галлюцинаций. В то же время отсутствие таких акцентов в кратких, но информационно насыщенных фрагментах приводит к пропускам или искаженному восприятию микротем. Некоторые модели при этом пытаются искусственно структурировать такие фрагменты, что часто сопровождается фактологическими ошибками.

2. Средства визуальной и структурной наглядности облегчают членение текста для LLMs.

Формулы, таблицы, графики, списки и иллюстрации служат эффективными структурными маркерами, которые позволяют языковым моделям легче выделять и классифицировать микротемы для последующего включения в реферат. Их отсутствие затрудняет внутреннюю сегментацию текста.

3. Большое количество разделов и подразделов усложняет задачу ранжирования микротем.

В случае объемных детализированных научных текстов LLMs часто стремятся объединять мелкие смысловые блоки с более крупными, а также перестраивать их порядок следования. Хотя такая трансформация не нарушает общей логики изложения, она может привести к выпадению значимых микротем.

4. Приоритет формальной структуры над содержанием.

При попытке сохранить оригинальную структуру научной статьи LLMs зачастую фокусируются на воспроизведении заголовков разделов, пренебрегая их смысловым наполнением. В результате в реферате могут присутствовать номинальные ссылки на подразделы без адекватной передачи их содержания.

5. Сильная зависимость от формальных признаков заголовков.

LLMs демонстрируют склонность к устранению фрагментов с семантически близкими названиями, ошибочно интерпретируя их как дублирующие. Это может указывать на наличие у моделей механизма семантической экономии, при котором схожие наименования воспринимаются как признак избыточности информации.

6. Имплицитные представления о канонической структуре научного текста.

Модели склонны опускать структурные элементы, которые не вписываются в предполагаемую логику изложения научной статьи. Так, например, подраздел с историографической справкой, расположенный между подразделами, посвященными предмету и методологии исследования, был исключен из реферата. Это подтверждает наличие у LLMs внутренних моделей «канонической» структуры научного текста. Примером может служить также затруднение в идентификации авторов, если информация о них расположена в нетипичном месте – например, в конце статьи.

7. Избыточно развернутая аннотация дезориентирует LLMs.

Если аннотация слишком подробна и дублирует содержание основной части, модели воспринимают ее как самостоятельный смысловой блок. Это приводит к игнорированию повторяющейся информации в статье и, как следствие, к потере важных микротем, ошибочно интерпретированных как уже обработанные.

8. Зависимость от наличия дискурсивных слов-маркеров.

LLMs используют лексические маркеры (такие как *цель, актуальность объект, предмет, задача, литература*) для определения значимости фрагментов текста. При отсутствии таких сигналов и необходимости контекстуального анализа вероятность исключения важной информации возрастает.

Исследование автоматических рефератов русскоязычных научных текстов показало, что качество работы LLMs существенно зависит от множества факторов – как лингвистических, так и структурных. Установленные зависимости позволяют более точно формулировать принципы подготовки научных текстов, ориентированных на машинную обработку. Выдвинутые гипотезы требуют дальнейшей верификации на расширенном корпусе текстов и могут лечь в основу предстоящих исследований, направленных на совершенствование реферативной функции больших языковых моделей при обработке русскоязычных научно-технических статей.

ЛИТЕРАТУРА

1. *Katsnelson, M.* Emergent Quantumness in Neural Networks / M. Katsnelson, V. Vanchurin // Foundation of Physics. – 2021. – Vol. 51, № 94. – P. 1–20.
2. *Кукель, В. Е.* Реальность алгоритма – интенциональность и инструментальность / В. Е. Кукель // Интеллект. Инновации. Инвестиции. – 2025. – № 3. – С. 141–148.
3. *Кузнецов, А. Г.* Туманности нейросетей: «черные ящики» технологий и наглядные уроки непрозрачности алгоритмов / А. Г. Кузнецов // Социология власти. – 2020. – № 32 (2). – С. 157–182.
4. *Васильевская, В. М.* Потенциальные возможности компьютерной технологии YaGPT для автоматического реферирования текста: содержательный аспект / В. М. Васильевская // Вестник МГЛУ. Сер. 1, Филология. – 2025. – № 2 (135). – С. 139–151.
5. *Васильевская, В. М.* Лингвистические ошибки нейросетей при реферировании русскоязычных текстов / В. М. Васильевская // От слова к дискурсу : материалы Междунар. науч. конф., Минск, 15–17 мая 2025 г. / Минск. гос. лингвист. ун-т ; редкол.: Ю. В. Овсейчик (отв. ред.) [и др.]. – Минск : МГЛУ, 2025. – С. 290–292.
6. *Леонов, В. П.* Реферирование и аннотирование научно-технической литературы // В. П. Леонов. – Новосибирск : Наука, 1986. – 172 с.
7. Научная электронная библиотека «КиберЛенинка» : [сайт]. – URL: https://cyberleninka.ru/ (дата обращения: 17.08.2025).

Поступила в редакцию 26.08.2025