ISSN: 2224-5227 (Print)
ISSN: 2518-1483 (Online)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№1
2026

ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

1 (357)

JANUARY – MARCH 2026

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: «Central Asian Academic Research Center» LLP (Almaty).
Certificate **№ KZ77VPY00121154** on the re-registration of the periodical printed and online publication of the information agency, issued on **05.06.2025** by the Republican State Institution «Information Committee» of the Ministry of Culture and Information of the Republic of Kazakhstan
Subject area*: information and communication technologies.*
Currently: *included in the list of journals recommended by the CCSES MSHE RK in the direction of «Information and communication technologies».*
Periodicity: *4 times a year.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Меншіктеуші: «Орталық Азия академиялық ғылыми орталығы» ЖШС (Алматы).
Ақпарат агенттігінің мерзімді баспасөз басылымын, ақпарат агенттігін және желілік басылымды қайта есепке қою туралы ҚР Мәдениет және Ақпарат министрлігі «Ақпарат комитеті» Республикалық мемлекеттік мекемесі **05.06.2025** ж. берген **№ KZ77VPY00121154** Куәлік.
Тақырыптық бағыты: *ақпараттық-коммуникациялық технологиялар*
Қазіргі уақытта: *«ақпараттық-коммуникациялық технологиялар» бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *ТОО «Центрально-азиатский академический научный центр» (г. Алматы).*
Свидетельство о постановке на переучет периодического печатного издания, информационного агентства и сетевого издания **№ KZ77VPY00121154.** Дата выдачи **05.06.2025**
Тематическая направленность: *информационно-коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных КОКСНВО МНВО РК по направлению «информационно-коммуникационные технологии».*
Периодичность: *4 раза в год.*
http://www.physico-mathematical.kz/index.php/en/

CONTENTS

COMPUTER SCIENCE

МАЗМҰНЫ

ИНФОРМАТИКА

СОДЕРЖАНИЕ

ИНФОРМАТИКА

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 316–336

https://doi.org/10.32014/2026.2518-1726.415

IRSTI 28.23.37
UDC 004.89

© Oralbekova D.[1*], Mamyrbayev O. [1], Akhmediyarova A.[2], Kassymova D. [3], Alibiyeva Z. [2], 2026.
[1] Institute of information and computational technologies, Almaty, Kazakhstan;
[2] Satbayev University, Almaty, Kazakhstan;
[3] Mukhametzhan Tynyshbayev ALT University, Almaty, Kazakhstan.
*E-mail: dinaoral@mail.ru

DEVELOPMENT OF A MULTI-LEVEL MODEL FOR TEXT SUMMARIZATION BASED ON PRETRAINED MODELS

Oralbekova Dina — PhD, associate professor, Institute of information and computational technologies, Almaty, Kazakhstan,
E-mail: dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493;
Mamyrbayev Orken — PhD, professor, Institute of information and computational technologies, Almaty, Kazakhstan,
E-mail: morkenj@mail.ru, https://orcid.org/0000-0001-8318-3794;
Akhmediyarova Ainur — PhD, professor, Satbayev University, Almaty, Kazakhstan,
E-mail: a.akhmediyarova@satbayev.university, https://orcid.org/0000-0003-4439-7313;
Kassymova Dinara — PhD, associate professor, Mukhametzhan Tynyshbayev ALT University, Almaty, Kazakhstan,
E-mail: d.kassymova@alt.edu.kz, https://orcid.org/0000-0001-6152-8317;
Alibiyeva Zhibek — PhD, associate professor, Satbayev University, Almaty, Kazakhstan,
E-mail: zh.alibiyeva@satbayev.university, ORCID ID: https://orcid.org/0000-0001-9565-5621.

Abstract. This paper explores the application of modern transformer models to the task of abstract text summarization in Kazakh, a resource-poor language characterized by an agglutinative structure and complex morphology. These language features complicate the application of traditional approaches to text processing, necessitating the use of more adapted models. The proposed summarization system utilizes multi-level text processing, including symbolic, subword, word, and contextual representation levels. This structure allows for consideration of both the morphological and semantic properties of the Kazakh language. The base models used are the multilingual transformers mBART, mT5, and XLM-RoBERTa, which were adapted and further trained for the task of abstract text summarization. A specialized corpus of 1,000 Kazakh-language news articles with manually compiled annotations was compiled for training and evaluating

the quality of the models. Preprocessing utilized character representations, SentencePiece word-by-word tokenization, FastText word vectors, and contextual transformer embeddings. The quality of the generated summaries was assessed using a set of automated metrics, including ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR, and BERTScore F1, allowing for the analysis of both superficial matches and semantic correspondence with reference annotations. Experimental results showed that the mBART model demonstrated the best performance across most metrics, while the combination of XLM-RoBERTa and BART also yielded stable and competitive results. The experiments demonstrated that a multi-level text processing scheme contributes to improved abstract summaries. Modern transformer models also demonstrate good results when working with Kazakh-language texts.

Keywords: automatic summarization, multi-level language modeling, Kazakh language, transformer architectures, hybrid approach

For citations: Oralbekova D., Mamyrbayev O., Akhmediyarova A., Kassymova D., Alibiyeva Z. Development of a multi-level model for text summarization based on pretrained models. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 316–336. DOI: https://doi.org/10.32014/2026.2518-1726.415

Оралбекова Д.[1*], Мамырбаев О.[1], Ахмедиярова А.[2], Қасымова Д.[3], Алибиева Ж.[2], 2026.

[1] Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан;

[2] Satbayev университеті, Алматы, Қазақстан;

[3] Mukhametzhan Tynyshbayev ALT University, Алматы, Қазақстан.

E-mail: dinaoral@mail.ru

АЛДЫН АЛА ОҚЫТЫЛҒАН МОДЕЛЬДЕР НЕГІЗІНДЕ МӘТІНДІ РЕЗЮМЕЛЕУГЕ АРНАЛҒАН КӨПДЕҢГЕЙЛІ МОДЕЛЬДІ ӘЗІРЛЕУ

Оралбекова Дина — PhD, қауымд. проф., Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан,
E-mail: dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493;
Мамырбаев Өркен — PhD, профессор, бас директордың орынбасары, Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан,
E-mail: morkenj@mail.ru, https://orcid.org/0000-0001-8318-3794;
Ахмедиярова Айнұр — PhD, профессор, Satbayev университеті, Алматы, Қазақстан,
E-mail: a.akhmediyarova@satbayev.university, https://orcid.org/0000-0003-4439-7313;
Қасымова Динара — PhD, қауымд. проф., Mukhametzhan Tynyshbayev ALT University, Алматы, Қазақстан,
E-mail: d.kassymova@alt.edu.kz, https://orcid.org/0000-0001-6152-8317.
Алибиева Жибек — PhD, қауымд. проф., Satbayev университеті, Алматы, Қазақстан,
E-mail: zh.alibiyeva@satbayev.university, ORCID ID: https://orcid.org/0000-0001-9565-5621.

Аннотация. Бұл мақалада агглютинативті құрылымы мен күрделі морфологиясымен сипатталатын ресурстары аз қазақ тіліндегі абстрактілі мәтінді қысқаша баяндау міндетіне заманауи трансформерлік модельдерді қолдану қарастырылады. Бұл тілдік ерекшеліктер мәтінді өңдеуге дәстүрлі тәсілдерді қолдануды қиындатады, ал бұл бейімделген модельдерді пайдалануды қажет етеді. Ұсынылған қысқаша резюмелеу жүйесі символдық, сөздік, және контекстік көрініс деңгейлерін қоса алғанда, көп деңгейлі мәтінді өңдеуді пайдаланады. Бұл құрылым қазақ тілінің морфологиялық және семантикалық қасиеттерін ескеруге мүмкіндік береді. Қолданылатын базалық модельдер абстрактілі мәтінді қысқаша баяндау міндетіне бейімделген және одан әрі оқытылған көптілді mBART, mT5 және XLM-RoBERTa трансформер модельдері болып табылады. Оқыту және бағалау үшін қолмен дайындалған эталондық аннотациялары бар қазақ тіліндегі 1 000 жаңалық мақаласынан тұратын арнайы корпус құрастырылды. Алдын ала өңдеу кезеңінде таңбалық деңгейдегі ұсынылымдар, SentencePiece негізіндегі субсөздік токенизация, FastText сөздік ендірулері және трансформер модельдерінен алынған контекстік ендірулер пайдаланылды. Жасалған қысқаша резюмелеу сапасы ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR және BERTScore F1 сияқты автоматтандырылған көрсеткіштер жиынтығын пайдаланып бағаланды, бұл үстірт сәйкестіктерді де, сілтеме аңдатпаларымен семантикалық сәйкестікті де талдауға мүмкіндік берді. Тәжірибелік нәтижелер mBART моделі көптеген көрсеткіштер бойынша ең жақсы өнімділікті көрсеткенін, ал XLM-RoBERTa және BART комбинациясы тұрақты және бәсекеге қабілетті нәтижелер бергенін көрсетті. Тәжірибелер көп деңгейлі мәтінді өңдеу схемасы абстрактілі қорытындыларды жақсартуға ықпал ететінін көрсетті. Қазіргі заманғы трансформаторлық модельдер қазақ тіліндегі мәтіндермен жұмыс істеген кезде де жақсы нәтижелер көрсетеді.

Түйін сөздер: автоматты резюмелеу, көп деңгейлі тілдік модельдеу, қазақ тілі, трансформерлік архитектуралар, гибридті тәсіл

Оралбекова Д.[1*], Мамырбаев О.[1], Ахмедиярова А.[2], Касымова Д.[3], Алибиева Ж.[2], 2026.

[1] Институт информационных и вычислительных технологий, Алматы, Казахстан;

[2] Satbayev University, Алматы, Казахстан;

[3] ALT университет имени М. Тынышпаева, Алматы, Казахстан.
E-mail: dinaoral@mail.ru

РАЗРАБОТКА МНОГОУРОВНЕВОЙ МОДЕЛИ В ЗАДАЧАХ РЕЗЮМИРОВАНИЯ ТЕКСТА НА ОСНОВЕ ПРЕДВАРИТЕЛЬНО ОБУЧЕННЫХ МОДЕЛЕЙ

Оралбекова Дина — PhD, ассоц. проф., Институт информационных и вычислительных технологий, Алматы, Казахстан,

E-mail: dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493;
Мамырбаев Оркен — PhD, профессор, заместитель генерального директора, Институт информационных и вычислительных технологий. Алматы, Казахстан,
E-mail: morkenj@mail.ru, https://orcid.org/0000-0001-8318-3794;
Ахмедиярова Айнур — PhD, профессор, Satbayev Университет, Алматы, Казахстан,
E-mail: a.akhmediyarova@satbayev.university, https://orcid.org/0000-0003-4439-7313;
Касымова Динара — PhD ассоц. проф., ALT университет имени М. Тынышпаева, Алматы, Казахстан,
E-mail: d.kassymova@alt.edu.kz, https://orcid.org/0000-0001-6152-8317;
Алибиева Жибек — PhD, ассоц. проф., Satbayev Университет, Алматы, Казахстан,
E-mail: zh.alibiyeva@satbayev.university, https://orcid.org/0000-0001-9565-5621.

Аннотация. В работе исследуется применение современных трансформерных моделей для задачи абстрактивного резюмирования текстов на казахском языке, относящемся к малоресурсным языкам и характеризующемся агглютинативной структурой и сложной морфологией. Эти особенности затрудняют использование традиционных методов обработки текста и требуют применения более адаптированных подходов. Предложенная система основана на многоуровневой обработке текста, включающей символьный, подсловный, словный и контекстный уровни представления. Такой подход позволяет учитывать как морфологические, так и семантические характеристики казахского языка. В качестве базовых моделей использованы многоязычные трансформеры mBART, mT5 и XLM-RoBERTa, которые были адаптированы и дообучены для задачи абстрактивного резюмирования. Для обучения и оценки качества моделей сформирован специализированный корпус, включающий 1000 новостных статей на казахском языке с вручную составленными аннотациями. В процессе предобработки применялись символьные представления, подсловная токенизация SentencePiece, словные векторные представления FastText, а также контекстные эмбеддинги трансформерных моделей. Оценка качества сгенерированных резюме проводилась с использованием набора автоматических метрик, включая ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR и BERTScore (F1), что позволило учесть как лексическое совпадение, так и семантическую близость к эталонным аннотациям. Экспериментальные результаты показали, что модель mBART демонстрирует наилучшие показатели по большинству метрик, тогда как комбинация XLM-RoBERTa и BART обеспечивает стабильные и конкурентоспособные результаты. Проведённые эксперименты подтверждают, что многоуровневая схема обработки текста способствует повышению качества абстрактивного резюмирования. Использование современных трансформерных моделей также демонстрирует высокую эффективность при работе с казахскоязычными текстами.

Ключевые слова: автоматическое резюмирование, многоуровневое языковое моделирование, казахский язык, трансформерные архитектуры, гибридный подход

Финансирование. *Данное исследование финансировалось Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант BR24993166).*

Введение. Языковое моделирование (ЯМ) представляет собой одну из центральных задач в области обработки естественного языка (NLP), которая включает в себя создание и развитие алгоритмов, способных обрабатывать, понимать и генерировать текстовые данные на человеческом языке. В последние годы с развитием алгоритмов глубокого обучения языковое моделирование стало одним из популярных направлений в NLP. Появление архитектур на основе Трансформера (Vaswani et al., 2017; Rahali et al., 2023) внесли существенный вклад в понимании контекста данных, структуру предложения и семантические взаимосвязи слов. На сегодняшний день такие модели используются при решении различных задач NLP, включая машинный перевод, автоматическое составление кратких аннотаций, системы не основе вопросов и ответом, а также в анализе тональности текстов (Narejo et al., 2024; Oralbekova et al., 2024).

Среди задач языкового моделирования значительное внимание уделяется автоматическому резюмированию текста. Существующие методы можно разделить на два типа. К первому можно отнести extractive summarization, который извлекает ключевые предложения из исходного текста. Вторым методом является abstractive summarization, при котором идет генерация нового текста на основе исходного контента (Giarelis et al., 2023). В последние годы модели трансформеров, такие как BERT (Kalyan et al., 2021), GPT (Liu et al., 2024), T5 (Raffel et al., 2020) и BART (Lewis et al., 2020), показали значительные успехи в задачах генерации текста, включая резюмирование. Для абстрактного резюмирования часто используются модели BART и T5. Их архитектура на основе кодер-декодер обеспечивают преобразования исходного текста в краткую и связную форму. Кроме традиционных методов резюмирования, есть еще гибридных подход, который сочетает характеристики обоих типов аннотирования (Kirmani et al., 2019). Гибридное резюмирование включает в себя извлечение ключевых фрагментов текста с последующей генерацией резюме на основе этих фрагментов, что позволяет сочетать точность extractive и гибкость генерации. Гибридные методы резюмирования объединяют свойства извлекательного и abstractive подходов. С их помощью можно сохранить ключевые элементы исходного текста и в тоже время сформировать компактное и связное изложение содеражания. Такой подход особенно актуален для казахского языка, потому что текстовые корпуса и цифровые контенты на этом языке остаются ограниченными.

Развитие многоязычных архитектур также сыграло важную роль в этой области. Модели mBART и mT5 обучаются на текстах разных языков. Они способны работать даже с языками, для которых объем обучающих данных относительно невелик. Именно это особенность позволяет применять общие

алгоритмические подходы при обработке многоязычных текстов. mBART (Multilingual BART) и mT5 (Multilingual T5) продемонстрировали высокие результаты в многозадачных сценариях, таких как перевод, анализ текста и генерация резюме. Модель XLM-RoBERTa (Conneau et al., 2020) представляет собой многоязычное расширение архитектуры RoBERTa (Liu et al., 2019). Данная модель показала высокую эффективность при решении различных задач NLP, включая классификацию текста и аннотирование информации.

Современные языковые модели достигли заметных успехов при обработке распространённых языков, включая английский, французский и китайский. Однако для малоресурсных языков подобные результаты пока достигнуты не в полной мере. К таким языкам можно отнести семейство агглютинативных и тюркоязычных языков, куда и входит казахский язык, что требуют использования специальных методов обработки текста. Кроме этого, объём доступных текстовых корпусов и количество предобученных моделей для казахского языка остаются ограниченными. Поэтому автоматическое резюмирование казахских текстов недостаточно исследованы, а известные подходы показывают умеренные показатели качества (Zulkhazhav et al., 2019; Kynabay et al., 2021; Zhabayev et al., 2021).

В настоящей работе предложена многоуровневая ЯМ, разработанная специально для казахского языка. Мы представляем модель, которая обрабатывает текст на четырех уровнях: символьном, подсловном, словном и контекстном. Применение такой структуры позволяет более полно учитывать морфологические характеристики языка и текста. Кроме того, данный подход способствует повышению точности и информативности автоматически формируемых аннотаций. В данном исследовании предлагается использование моделей mT5, mBART и XLM-RoBERTa, которые были адаптированы и дообучены для казахского языка.

В разделе 2 представлена обзорная литература по задаче резюмирования текста с акцентом на существующие методы и модели, включая как экстрактивные, так и абстрактивные подходы, а также гибридные методы. В разделе 3 подробно описаны используемые модели и их архитектуры, включая многоязычные трансформеры, такие как mT5, mBART и XLM-RoBERTa, а также принципы их адаптации для задачи резюмирования на казахском языке. В разделе 4 изложены методы проведения экспериментов и полученные результаты, с анализом их значимости и обсуждением достигнутых показателей. Раздел 5 содержит заключение, в котором подведены итоги работы, обсуждаются направления для будущих исследований в области NLP.

Материалы и методы исследования. В последние годы значительные усилия были направлены на разработку гибридных моделей резюмирования, ориентированных как на ресурсообеспеченные, так и на низкоресурсные языки. В последние годы опубликован ряд работ, посвящённых разработке моделей автоматического резюмирования для малоресурсных языков.

В работе (Winarko et al., 2025) предложили модель абстрактного резюмирования для индонезийского языка. В данной архитектуре в качестве энкодера используется стек эмбеддингов, а декодирующая часть построена на Трансформере. В качестве эмбеддингов использовались BERT, Byte Pair Encoding (BPE), Character Embedding (CE) и FastText. Авторы провели исследование влияния выбора слоёв BERT и конфигурации эмбеддингов на качество резюме. Модель обучалась на подкорпусах Liputan6 (50K и 75K статей). Наиболее высокие значения ROUGE-1 (37.18), ROUGE-2 (18.19) и ROUGE-L (34.28) были достигнуты при использовании всех слоёв BERT. Предложенная архитектура показала высокие результаты даже при ограниченном обучающем объёме данных.

В исследовании (Raza et al., 2024) описана гибридная система резюмирования текстов на языке урду. Авторы объединили два традиционных метода резюмирования. На первом этапе применялись классические подходы, основанные на TF-IDF, весах предложений и частоте употребления слов Полученный фрагмент затем использовался в качестве входных данных для абстрактной модели, построенной на архитектуре BERT. Качество сформированных резюме оценивалось специалистами, владеющими языком урду. Для обучения и тестирования был сформирован корпус из материалов изданий Express, BBC Urdu, Dawn и других источников на различные темы. Данное исследование относится к числу ранних работ по абстрактному резюмированию текстов на урду и учитывает морфологическую сложность и лексическое разнообразие языка.

В работе (Challagundla et al., 2024) была разработана комплексная структура модели для абстрактного резюмирования. Предложенный подход объединил методы seq2seq с механизмом внимания, семантическое обобщение и разрешении омонимии. Векторизация текста выполнялась с помощью Word2Vec, архитектура включала двунаправленные LSTM, кастомный attention и TimeDistributed слой. Для обучения и тестирования модели использовался корпус из материалов Gigaword, DUC-2004 и CNN/DailyMail. Использование WSD и постобработки улучшило согласованность и охват редких слов, повысив качество резюме по сравнению с базовыми моделями seq2seq без семантических дополнений.

Для обработки патентных документов был предложил гибридный подход к резюмированию (Jayatilleke et al. 2025). В нём сочетается экстрактивный алгоритм LexRank и модель BART, дополнительно адаптированный с применением метода Low-Rank Adaptation (LoRA). Модель дополнительно обобщалась на новые патентные области с помощью методов метаобучения. Архитектура была протестирована на длинных патентных текстах, отличающихся сложной юридической терминологией. Благодаря LoRA удалось сократить вычислительные ресурсы при сохранении качества. Такой подход оказался эффективным для создания абстрактных резюме в области интеллектуальной собственности.

В работе (Do et al., 2021) предложили Discrete Diffusion Language Model (DDLM) с семантически управляемым шумом (Semantic-Aware Noising) и архитектурой CrossMamba — адаптацией модели Mamba для задач резюмирования длинных текстов. Их модель показала превосходство над ранее предложенными дискретными диффузионными моделями и даже опередила автопорождённые (autoregressive) модели по скорости генерации на наборах данных CNN/DailyMail, Arxiv и Gigaword. В данной работе предлагается заменить случайный шум семантически ориентированным. Это позволяет сохранять важные токены и контролировать структуру формируемого текста.

В (Gogireddy et al., 2024) рассматривается система резюмирования, состоящая из двух этапов. На первом этапе используется Graph Neural Networks для выбора наиболее значимых дискурсивных единиц. Затем полученные данные передаются в модель BART для формирования окончательного варианта аннотации. Эксперименты проводились на CNN/DailyMail, и предложенный подход показал более высокие значения ROUGE по сравнению с базовыми трансформерами. Метод демонстрирует эффективность объединения графовых структур с генеративными трансформерами.

Исследование (Faizal et al., 2024) посвящено задаче абстрактного резюмирования финансовых и бизнес-отчётов. В предложенном подходе объединяется анализ текстовых и табличных данных. Модель построена на основе архитектуры Transformer и дополнена элементами Switch Transformer для эффективной обработки многоформатного финансового контента. Особенностью работы является адаптация модели к структурам таблиц, характерным для финансовых документов. Эксперименты проводились на бизнес-датасетах Reuters Financial и Bloomberg. Качество модели оценивалось с помощью метрик ROUGE и F-measure, показавших улучшенные результаты по сравнению с классическим Transformer-ом.

В (Rounak Chakraborti et al., 2025) провели сравнительный анализ различных моделей резюмирования, включая RNN, LSTM, Seq2Seq, BART, T5 и Pegasus. Исследование сосредоточено на задачах новостного резюмирования, где анализировалась точность, связность и общая пригодность моделей. Было использовано несколько корпусов, включая CNN/DailyMail и BBC News. В работе отмечены преимущества моделей Pegasus и BART, особенно в части генерации контекстуально осмысленных резюме. Также обсуждаются ограничения метрик ROUGE и необходимость более сложных оценок (например, по связности и достоверности).

В (Rao et al., 2025) представили новую метрику оценки для абстрактного резюмирования, основанную на семантическом сходстве между референсным и сгенерированным резюме. В качестве базовой модели использовался Universal Sentence Encoder (USE), позволяющий сравнивать смыслы на уровне предложений. Было показано, что предложенная метрика

лучше коррелирует с оценками людей, чем ROUGE, особенно при низком n-граммовом совпадении. Эксперименты проводились на CNN/DailyMail и других наборах данных. В работе также рассмотрены другие альтернативные метрики: BERTScore, MoverScore и ROUGE-G.

В исследовательской работе (Zangooei et al., 2025) разработали модель ARLED для абстрактного резюмирования длинных текстов на персидском языке. Архитектура объединила модели ARMAN (на основе Longformer) и LED (Longformer Encoder-Decoder), что позволило эффективно работать с входами до 8192 токенов. Был создан новый корпус из 49,457 персидских научных статей с Ensani.ir, прошедший глубокую предобработку. Токенизация данных выполнялась с помощью HuggingFace AutoTokenizer. Кроме того, в процессе обучения были реализованы стратегии семантической переупорядоченности предложений. Подобный подход ранее редко применялся при работе с персидским языком.

В (Langston et al., 2024) рассматривается автоматизированный подход к резюмированию аннотаций и заголовков нескольких документов с использованием крупных языковых моделей (LLMs). Предложенная система использует комбинированную стратегию. На первом этапе из текста извлекаются ключевые фрагменты, после чего выполняется их абстрактное обобщение при помощи ЯМ. Эксперименты показали, что система демонстрирует особенно высокую эффективность при работе с техническими и медицинскими текстами. В работе также отмечается важность адаптации моделей к конкретной предметной области и рассматриваются перспективы использования методов обучения с подкреплением (RL) и специализированных механизмов внимания.

Авторы (Barta et al., 2024) представили открытый венгерский корпус HunSum-2, сформированный на основе новостных текстов из Common Crawl и предназначенный для задач как экстрактивного, так и абстрактивного резюмирования. Для предварительной обработки данных применялись FastText (для определения языка) и quntoken (для токенизации и разбиения на предложения). В абстрактивных экспериментах использовались модели на базе mT5 и архитектура Bert2Bert с ограничением на длину входной последовательности до 512 токенов. В экстрактивном подходе были реализованы модели BertSum с использованием huBERT и простого классификатора, где метки предложений генерировались с опорой на косинусное расстояние между их эмбеддингами. Анализ экспериментальных данных показал, что по метрикам ROUGE и BERTScore более высокие значения были получены для экстрактивных моделей. Среди абстрактивных систем наиболее конкурентоспособной оказалась модель mT5, однако её показатели всё же уступали экстрактивным методам. При этом абстрактивные модели чаще допускали несоответствия и отклонения от исходного содержания.

Алгоритм многоуровневого моделирования

Многоуровневое моделирование основывается на идее о том, что для

успешной обработки текста, особенно для языков с богатой морфологией и сложными структурами, необходимо учитывать различные уровни текстовых единиц. В разработанной архитектуре предусмотрено несколько уровней обработки данных (алгоритм 1):

- Символьный уровень предназначен для анализа структуры слова. Для казахского языка такой подход особенно важен, поскольку изменения в морфемах могут существенно менять смысл слова. В модели применяются представления character-level embeddings для обеспечения обработки текста на уровне отдельных символов (Jebbara et al., 2017).

- При работе с агглютинативными языками, такими как казахский, важным этапом является подсловная обработка текста. В таких языках одно слово может включать несколько морфем и иметь сложную структуру. Использование подхода вроде Sentence Piece позволяет модель обрабатывать текст на уровне частей слов (Kudo et al., 2018).

- Словный уровень помогает выявить семантическую составляющую текста. Применение векторных представлений слов с помощью FastText позволяет модели понимать семантическое значение отдельных слов (Stein et al., 2019).

- На контекстном уровне модель анализирует взаимосвязи между словами и предложениями. Архитектуры Transformer дают возможность учитывать контекст их употребления, и грамматически корректно генерировать краткое содержание текста.

Алгоритм 1: Гибридный подход к резюмированию текста на основе предобученных моделей

Цель: Алгоритм предназначен для генерации кратких резюме казахских текстов. Предлагаемый подход основан на многоуровневой обработке данных, включающей символьный, подсловный, словный и контекстный уровни представления текста.

Входные данные: Корпус казахских текстов с соответствующими резюме.

Выходные данные: Сгенерированные резюме для наиболее представительных казахских текстов.

– Производится загрузка датасета, включающего казахские тексты и подготовленные к ним резюме.

– Осуществляется предобработка входных данных.

– К входным данным добавляется специальный префикс «summarize:»

– Токенизация текста на символьном уровне с использованием моделей, таких как Character-level embeddings, чтобы захватить структуру слов, особенно для казахского языка, где морфологические изменения критичны.

– Обработка текста включает этап подсловной токенизации, для которого применяются алгоритмы Byte Pair Encoding (BPE) или SentencePiece.

– Аналогичная процедура токенизации применяется и к соответствующим резюме.

1. Загрузка и инициализация предобученной модели, применяемой для задач типа seq2seq.

2. Дообучение модели на подготовленном датасете (казахские тексты с резюме) с использованием соответствующих параметров обучения, таких как скорость обучения, размер батча и количество эпох.

3. В процессе обучения выполняется вычисление оценочных метрик, которые отслеживают динамику качества работы модели.

4. Для каждого входного текста обученная модель формирует краткое резюме. При этом учитываются контекстные зависимости между словами на уровне всего текста.

5. На завершающем этапе выводятся сгенерированные резюме, после чего проводится оценка качества модели на основе полученных значений метрик.

Модель mT5

mT5 представляет собой многоязычную модификацию модели T5, обученную на корпусах более чем 100 языков, включая казахский. Архитектура модели построена по схеме кодер-декодер и ориентирована на решение задач обработки естественного языка в формате «от текста к тексту». Благодаря этому модель может применяться для различных задач, включая машинный перевод, генерацию текста и автоматическое резюмирование.

Архитектурно mT5 основана на трансформере, где кодер обрабатывает входную последовательность, а декодер формирует выходной текст. Такая структура делает модель удобной для задач seq2seq, где требуется не только извлечение ключевых фрагментов, но и формирование нового текста из исходного документа. Модель mT5 использует SentencePiece токенизацию для преобразования текста в последовательности токенов. В архитектуре в Encoder входной текст разбивается на токены и передается через несколько слоев self-attention механизма для вычисления скрытых состояний. Каждый слой состоит из multi-head attention и feed-forward neural network. После прохождения всех слоев, выходные представления становятся контекстуальными эмбеддингами, которые затем передаются в декодер. В архитектуре модели декодер получает скрытые состояния, сформированные кодером и использует их для последовательной генерации текста. Механизм внимания позволяет учитывать ранее сгенерированные токены. В результате предсказание нового токена выполняется на основе уже сформированной части текста и внутренних представлений модели.

Модель mBART

mBART является многоязычной версией архитектуры BART. Данная модель применяется в задачах генерации текста, включая машинный перевод, суммирование и другие задачи, связанные с обработкой текстовых последовательностей. Модель mBART также как и mT5 использует архитектуру encoder-decoder (рис. 1).

mBART обучена на нескольких языках с использованием многоязычного корпуса. Это позволяет ей работать с текстами на различных языках, включая казахский. mBART использует SentencePiece для токенизации. Модель

продемонстрировала отличные результаты в задачах перевода, генерации и резюмирования текста, особенно в случае, когда нужно работать с текстами на малоресурсных языках. Архитектура mBART объединяет свойства моделей BERT и GPT. Двунаправленный механизм обработки контекста реализован в энкодере, тогда как декодер работает в авторегрессионном режиме.

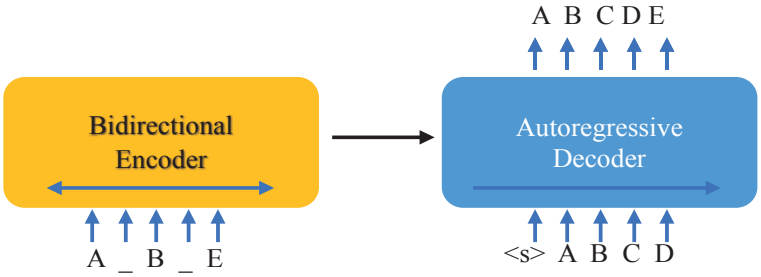

Рисунок 1 – BART architecture

Модель XLM-RoBERTa

XLM-RoBERTa (Cross-lingual RoBERTa) является улучшенной версией модели RoBERTa, разработанной для работы с многоязычными текстами. XLM-RoBERTa построена на основе архитектуры Transformer и обучена на масштабном многоязычном корпусе, содержащем тексты более чем на 100 языках, включая казахский. Архитектура модели предполагает использование двунаправленного контекста, что позволяет учитывать слова, расположенные по обе стороны от текущего токена. Такой механизм способствует более точному моделированию зависимостей в тексте. Благодаря обучению на многоязычных данных модель может эффективно применяться для малоресурсных языков. Она формирует контекстные представления слов и выявляет связи между ними даже в случае редких или морфологически сложных форм. Предобучение выполнялось с использованием стратегии маскированного предсказания токенов, что улучшает качество языковых представлений. В архитектуре модели кодер отвечает за обработку входной последовательности. В ходе кодирования токены преобразуются в контекстные эмбеддинги, отражающие их синтаксические и семантические связи. Применение механизма внимания позволяет учитывать окружение каждого токена с обеих сторон. После обработки формируется набор скрытых представлений, описывающих текст на более абстрактном уровне. Эти представления используются для извлечения признаков на словном и контекстном уровнях и служат основой для дальнейшего анализа и генерации текста.

Токенизация в модели XLM-RoBERTa выполняется с использованием алгоритма BPE. Этот метод разделяет слова на подсловные единицы и облегчает обработку редких или сложных слов и повышает устойчивость модели при работе с новыми языковыми данными.

Применение многоуровневой языковой модели в задаче автоматического резюмирования предполагает анализ текста на нескольких уровнях представления. Такой подход особенно актуален для языков с развитой морфологией. Каждый уровень архитектуры участвует в извлечении и обработке различных типов информации из текста. В рамках исследования используется модель, включающая символьный, подсловный, словный и контекстный уровни обработки.

Обработка текста на символьном уровне реализована с использованием character-level embeddings для учитывания внутреннюю структуру слов. Такой способ представления данных полезен для казахского языка, где словоформы могут значительно изменяться в зависимости от грамматических категорий, а одна и та же морфема иногда имеет несколько вариантов записи. Учет символьной информации облегчает обработку морфологических изменений и разнообразных словоформ.

На подсловном уровне применяется токенизация с использованием алгоритма SentencePiece для улавливания особенности аффиксации, характерные для казахского языка. Это позволяет обрабатывать составные слова и аффиксальные формы, которые часто встречаются в тексте на казахском языке. SentencePiece разбивает слова на более мелкие единицы (подслова). Это дает модели эффективно справляться с неизвестными словами и морфемами.

На словном уровне используется FastText для векторизации слов. Применение FastText позволяет учитывать семантические связи между словами и формировать контекстно-зависимые векторные представления. Одним из преимуществ данной модели является возможность построения представлений даже для слов, которые отсутствовали в обучающем корпусе. Для казахского языка это особенно актуально, поскольку многие слова отличаются сложной морфологической структурой или встречаются редко.

Для формирования контекстных признаков на данном уровне используется модель XLM-RoBERTa или другие трансформерные архитектуры. Их применение позволяет анализировать зависимости между словами внутри предложения и между предложениями в тексте. Модель XLM-RoBERTa способна учитывать как локальные, так и долгосрочные контекстные зависимости. Благодаря этому формируется более точное представление смысла предложения и условий употребления каждого слова. Полученные представления используются при генерации резюме, где важно не только выделить основную информацию, но и передать её в краткой форме.

После завершения многоуровневой обработки текста генерация резюме выполняется моделью BART. BART будет использовать контекстные признаки, полученные на основе представлений XLM-RoBERTa. BART использует архитектуру encoder-decoder, где энкодер извлекает информацию из текста, а декодер создает новое резюме, перефразируя основные идеи исходного текста (рис. 2). Это позволяет получать высококачественные

резюме, которые сохраняют суть исходного материала, но представляют его в более краткой и сжато изложенной форме.

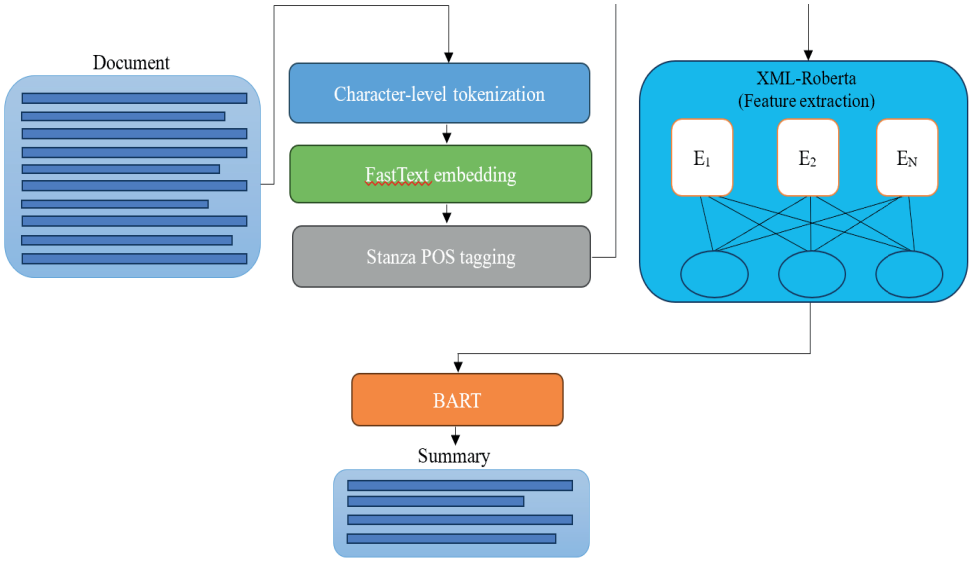

Рисунок 2 – Структура предложенного гибридного подхода

Для моделей mBART и mT5 дополнительное использование BART не требуется. Оба эти подхода уже имеют встроенную архитектуру encoder-decoder. Таким образом, модель mBART и mT5 могут непосредственно генерировать резюме. Эти архитектуры изначально построены по схеме seq2seq и предназначены для задач генерации текста.

Модель XLM-RoBERTa, напротив, реализует только encoder-часть трансформера. Она предназначена для извлечения признаков и не выполняет генерацию текста. По этой причине для формирования резюме используется BART, который выполняет функцию декодера и генерирует итоговое текстовое представление.

Результаты. Для обучения и тонкой настройки моделей был подготовлен корпус новостных материалов. Датасет включает новостные сообщения и аналитические статьи. Каждому тексту сопоставлена краткая аннотация, отражающая его основное содержание.Аннотации формировались вручную специалистом-лингвистом. Всего было собрано 1000 статей для обучения модели.

Для повышения согласованности и стабильности обучения, тексты были приведены к унифицированному формату: подавляющее большинство статей содержит не более 250 слов, а соответствующие аннотации — в среднем около 40 слов. Подобная организация данных позволила сбалансировать корпус по длине текстов и избежать существенных различий между короткими и длинными материалами.

Средняя длина статьи составила 112 слов, а медианная длина — 98 слов; длина резюме варьировалась в пределах 5–50 слов, со средним значением около 21 слова.

Предобработка данных

Перед обучением модели данные были подготовлены с использованием нескольких этапов предобработки. На этапе нормализации текст переводился в нижний регистр, из него удалялись лишние пробелы, а слова приводились к стандартной форме, например к их начальной форме. Это помогло улучшить качество данных и повысить эффективность модели. Из текста были удалены неинформативные символы, такие как специальные знаки, излишние пробелы, а также другие элементы, которые не вносят смысла в задачу резюмирования, такие как ссылки, HTML-теги.

Обработка текста на символьном уровне реализована с применением RNN-модели, которая позволяет учитывать информацию на уровне отдельных символов. Векторные представления слов формировались с помощью FastText. При обучении модели использовались следующие настройки: размер векторного пространства 300, окно контекста 5, а число эпох обучения составляло 10. После этапа токенизации выполнялось POS-размечивание с использованием библиотеки Stanza, и только затем данные передавались в модель для последующего извлечения признаков.

На этапе подготовки данных применялись разные стратегии токенизации в зависимости от архитектуры модели. Для mT5 и mBART использовался токенизатор SentencePiece, который формирует подсловные представления и облегчает обработку редких или составных слов. Для XLM-RoBERTa использовался стандартный токенизатор данной модели, основанный на алгоритме Byte Pair Encoding.

Размер словаря для mT5 и mBART был установлен на уровне 32 000 токенов, тогда как для XLM-RoBERTa он составлял 250 000 подслов. Модель mT5 была настроена с использованием 12 слоев и 768 размером скрытого состояния. Модель mBART использовала 12 слоев и 1024 размер скрытого состояния. Для XLM-RoBERTa и BART использовались 12 слоев с размером скрытого состояния 1024. Для обучения использовались 10 эпох с размером батча 16.

Для обучения использовалось 90% от всего корпуса, оставшиеся 10% были выделены для валидации. Все этапы реализации выполнялись с использованием библиотеки HuggingFace Transformers, включая модули datasets, tokenizers и Trainer, а также gensim для FastText и stanza для POS-аннотации.

При тонкой настройке модели применялись стандартные рекомендации для обучения трансформерных архитектур в задачах генерации текста. В качестве оптимизатора использовался AdamW со скоростью обучения 3e-5. Регуляризация обеспечивалась за счёт Dropout и weight decay с коэффициентом 0.01. Для стабилизации процесса обучения первые 10% шагов

от общего числа использовались в качестве этапа warmup. Чтобы увеличить эффективный размер батча, использовалось накопление градиентов. Норма градиента ограничивалась значением 1.0, а скорость обучения изменялась по линейному расписанию после этапа разогрева. Кроме того, применялся механизм early stopping. Обучение завершалось в случае отсутствия улучшения метрики eval_loss на валидационной выборке в течение трёх последовательных эпох.

Качество работы моделей резюмирования оценивалось с использованием нескольких метрик:

1) Метрики семейства ROUGE используются для сравнения сгенерированного текста с эталонным резюме. Оценка основана на совпадениях n-грамм, подстрок и последовательностей слов. Метрика ROUGE-1 фиксирует совпадения отдельных слов, ROUGE-2 учитывает совпадения пар слов, а ROUGE-L рассчитывается на основе наибольшей общей подпоследовательности. Каждая из метрик ROUGE вычисляется по формуле (1):

$$ROUGE - N = \frac{\sum_{overlap\ of\ N-grams} matches}{\sum_{N-grams\ in\ reference}} \quad (1)$$

где N - размер n-грамм, а «matches» - количество совпавших фрагментов между сгенерированным и эталонным резюме.

2) BLEU (Bilingual Evaluation Understudy) оценивает точность (precision) между сгенерированным и референсным текстом, сравнивая совпадения по n-граммам до 4-го порядка (2):

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} \omega_n \log\ p_n) \quad (2)$$

где - доля совпадающих n-грамм, ω_n - вес (обычно равный), а BP - penalty за слишком короткие предложения.

3) METEOR (Metric for Evaluation of Translation with Explicit ORdering) учитывает совпадения по словам, синонимам, стеммам и парафразам. По сравнению с другими метриками оценки качества резюмирования показатель METEOR более чувствителен к семантическому сходству. Формально (3):

$$METEOR = F_{mean} \cdot (1 - Penalty) \quad (3)$$

где F_{mean} - гармоническое среднее precision и recall, а Penalty зависит от количества перестановок.

4) BERTScore это метрика, основанная на контекстных эмбеддингах. Вместо подсчета поверхностных совпадений между токенами, BERTScore измеряет семантическое сходство между векторными представлениями слов,

извлечённых из предобученной модели. Значение метрики определяется как среднее максимальное косинусное сходство между токенами (4):

$$BERTScore\ F1\ = \frac{1}{N}\sum_{i=1}^{N}\max_{j}\ cos(\mathrm{v}_i, \mathrm{v}_j) \quad (4)$$

здесь N обозначает число токенов в сгенерированном тексте, v_i соответствует векторному представлению i-го токена сгенерированного текста, v_j – вектору j-го токена референсного текста. Функция $cos(\mathrm{v}_i, \mathrm{v}_j)$ определяет сходство между соответствующими векторами.

Данная метрика подходит для оценки качества резюме в случаях, когда сгенерированный текст передает смысл исходного, но не воспроизводит его дословно.

Проведение экспериментов и полученные результаты

Для сравнение полученных результатов были рассмотрены несколько работ для задачи резюмирования на казахских текстах. В настоящее время количество исследований, посвящённых автоматическому резюмированию казахских текстов, остаётся ограниченным. Большинство существующих подходов опираются на экстрактивные методы, такие как TF-IDF и логика нечёткого вывода. Работы по гибридному резюмированию появляются сравнительно недавно, в основном с применением трансформерных архитектур и стратегий трансферного обучения (табл. 1).

В работе (Zulkhazhav et al., 2019) использовалась экстрактивная модель на основе нечёткой логики с предварительным разрешением местоимений и морфологическим анализом. В ходе экспериментов для рассматриваемого метода были получены значения ROUGE-1 F1 = 0.44 и ROUGE-2 = 0.34. Для классического подхода без генеративной архитектуры такие результаты можно считать достаточно высокими.

Экстрактивный метод на основе TF-IDF также представлен в исследовании (Kynabay et al., 2021). В качестве источника текстов использовались новости портала inform.kz. Дополнительное использование специально сформированного списка стоп-слов для казахского языка позволило достичь ROUGE-L F1 = 0.35 без применения абстрактивных моделей.

В работе (Zhabayev et al., 2021) используется подход, при котором предложения сначала извлекаются на основе TF-IDF, после чего упрощаются с помощью модели Seq2Seq. Для обучения модели применялся подход трансферного обучения на корпусе Simple English Wikipedia. По результатам экспериментов значение метрики BLEU достигло 7.0, а показатели ROUGE не приведены.

Таблица 1. Результаты метрик для разных моделей

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BERTScore F1	Corpus Size
Zulkhazhav et al. (Fuzzy Logic, with pronoun resolution)	0,44	0,34	0,4	-	-	-	100 новостных текстов
Kynabay et al. (Extractive TF-IDF)	0,4	0,31	0,35	-	-	-	1422 новостных текстов
Zhabayev et al. (TF-IDF + Seq2Seq via Transfer Learning)	-	-	-	7	-	-	Корпус Kaz-skaz
mBART с многоуровневой архитектурой	0,59	0,41	0,51	0,49	0,52	91,40	1000 новостных текстов
mT5-small с многоуровневой архитектурой	0,52	0,37	0,48	0,32	0,37	91,20	1000 новостных текстов
XLM-R + BART с многоуровневой архитектурой	0,57	0,34	0,47	0,33	0,38	90,50	1000 новостных текстов

Обсуждение. На основе проведённых экспериментов и полученных метрик можно выделить несколько ключевых наблюдений, которые показывают различия в производительности моделей mBART, mT5 и XLM-RoBERTa. mBART продемонстрировала значительные улучшения по метрике ROUGE-1, достигая 0.5992. Полученные результаты свидетельствуют о высокой эффективности модели при формировании гибридных резюме. Дополнительным подтверждением служит постепенное увеличение значений метрик в ходе обучения.

Модель mT5 также демонстрирует положительную динамику по метрикам ROUGE-1 и ROUGE-L, достигая максимальных значений 0.5220 и 0.4840 соответственно. Однако её показатели остаются ниже результатов mBART, особенно по метрике ROUGE-2, где значение составляет 0.3760. Эти результаты, хотя и достойны внимания, всё же уступают показателям mBART и XLM-R + BART. Это свидетельствует о меньшей способности mT5 генерировать качественные резюме по сравнению другими моделями.

XLM-R + BART показала уверенное улучшение по метрике ROUGE-2, с максимальным значением 0.3440, а также по метрике ROUGE-L, с результатом 0.4790. Хотя модель XLM-R + BART продемонстрировала заметное улучшение показателей на последних этапах обучения, её значения ROUGE-1 и ROUGE-2 остаются ниже результатов других моделей. Тем не менее система сохраняет способность эффективно учитывать контекст и формировать достаточно точные резюме.

По значениям метрики METEOR модель mBART также демонстрирует наиболее высокие показатели при генерации гибридных резюме. В

экспериментах значение данной метрики достигало 0.5280. Это указывает на способность модели учитывать не только прямые совпадения слов, но и семантические соответствия, включая синонимы, стеммы и парафразированные выражения. Для mT5 и XLM-R + BART результаты по METEOR были ниже, но все равно показывали улучшения по сравнению с предыдущими исследованиями: mT5 достигла 0.3700, а XLM-R + BART — 0.3750. Полученные результаты подтвердили, что данные модели способны генерировать тексты с высоким уровнем семантического сходства.

Оценка по метрике BERTScore F1, предназначенной для измерения семантического сходства между сгенерированными и эталонными текстами, показывает, что модель mBART достигает наиболее высоких значений (0.9140). Такой результат указывает на способность модели сохранять смысловое содержание текста и корректно учитывать контекст при формировании резюме. Это особенно актуально для гибридных методов резюмирования, где ключевым требованием является точная передача смысла исходного документа.

Сравнительный анализ показал, что рассмотренные модели (mBART, mT5, XLM-R + BART) превосходят результаты, представленные в работах [14–16]. В указанных исследованиях значения метрики ROUGE-1 не превышают 0.44 (Zulkhazhav et al.) и 0.35 для ROUGE-L (для Kynabay et al.). Эти данные подтверждают, что современные трансформерные модели, такие как mBART, mT5 и XLM-R + BART, значительно улучшили производительность по сравнению с классическими экстрактивными методами, используемыми в предыдущих исследованиях.

Заключение. В данной работе была разработана и протестирована многоуровневая модель для гибридного резюмирования текстов на казахском языке с использованием современных трансформерных архитектур, таких как mBART, mT5 и XLM-RoBERTa. В представленном исследовании применяется многоуровневый подход к обработке текста. Использование нескольких уровней представления позволяет более полно учитывать особенности казахского языка, включая его морфологическую структуру и семантические связи. Именно такой гибридный подход способствует повышению качества создаваемых резюме. Результаты экспериментов показали, что модели mBART и XLM-RoBERTa, с использованием многоуровневой архитектуры, достигли наилучших результатов по метрикам ROUGE-1, ROUGE-2, ROUGE-L и BERTScore F1.

Дальнейшие исследования могут быть направлены на оптимизацию архитектуры модели, расширение корпуса обучающих данных и адаптацию системы к текстам разных жанров и стилевых типов.

References

Barta B., Lakatos D., Nagy A., Nyist M. K., Ács J. (2024) From News to Summaries: Building a Hungarian Corpus for Extractive and Abstractive Summarization. In Proceedings of the 2024

Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — P. 7503–7509. Torino, Italia. ELRA and ICCL (in Eng.)

Chakraborti R., Banerjee R., Das S. (2025) Evaluating the Efficacy of Text Summarization Models: A Comparison of NLP Algorithms. 2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, 2025. — P. 1-5, doi: 10.1109/IEMENTech65115.2025.10959463 (in Eng.)

Challagundla B. C., Peddavenkatagari C. (2024). Neural Sequence-to-Sequence Modeling with Attention by Leveraging Deep Learning Architectures for Enhanced Contextual Understanding in Abstractive Text Summarization. International Journal of Machine Learning and Cybernetics. 2. — P. 21-29 (in Eng.)

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. (2020) Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — P. 8440–8451 (in Eng.)

Do H. D., Duc A. D., Anh T. L., Wray B. (2025) Discrete Diffusion Language Model for Efficient Text Summarization. In Findings of the Association for Computational Linguistics: NAACL 2025. — P. 6278–6290 Albuquerque, New Mexico. Association for Computational Linguistics (in Eng.)

Faizal B., Abraham S., Thomas, S. (2024) Automated Business Report Summarization Using Transformer Model. 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), 1. — P. 254-258 (in Eng.)

Giarelis N.; Mastrokostas C.; Karacapilidis N. (2023) Abstractive vs. Extractive Summarization: An Experimental Review. Appl. Sci. 2023, 13. — P. 7620. https://doi.org/10.3390/app13137620 (in Eng.)

Gogireddy Y. R., Bandaru A. N., Sumanth V. (2024) Synergy of Graph-Based Sentence Selection and Transformer Fusion Techniques for Enhanced Text Summarization Performance. Journal of Computer Engineering and Technology (JCET) 7(1). 2024. — P. 33-41 (in Eng.)

Jayatilleke N., Weerasinghe Ruvan. (2025) A Hybrid Architecture with Efficient Fine Tuning for Abstractive Patent Document Summarization. 10.48550/arXiv.2503.10354 (in Eng.)

Jebbara S., Cimiano P. (2017) Improving Opinion-Target Extraction with Character-Level Word Embeddings. In Proceedings of the First Workshop on Subword and Character Level Models in NLP. — P. 159–167 (in Eng.)

Kalyan K.S., Rajasekharan A., Sangeetha S. (2021). AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing. ArXiv, abs/2108.05542 (in Eng.)

Kirmani M., Hakak N., Mohd, M., Mohd, M. (2019) Hybrid Text Summarization: A Survey: Proceedings of SoCTA 2017. 10.1007/978-981-13-0589-4_7 (in Eng.)

Kudo T., Richardson J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. — P. 66–71, Brussels, Belgium (in Eng.)

Kynabay B., Aldabergen A., Zhamanov A. (2021) Automatic Summarizing the News from Inform.kz by Using Natural Language Processing Tools. 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), Nur-Sultan, Kazakhstan, 2021. — P. 1-4, doi: 10.1109/SIST50301.2021.9465885 (in Eng.)

Langston O., Ashford B. (2024) Automated Summarization of Multiple Document Abstracts and Contents Using Large Language Models. TechRxiv. August 02, 2024. DOI: 10.36227/techrxiv.172262754.45577350/v1 (in Eng.)

Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. (2020) BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — P. 7871–7880 (in Eng.)

Liu X., Zheng Y., Du Z., Ding M., Qian Y., Yang Z., Tang J. GPT understands, too. AI Open. Vol. 5, 2024. — P. 208-215. ISSN 2666-6510, https://doi.org/10.1016/j.aiopen.2023.08.012 (in Eng.)

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692 (in Eng.)

Oralbekova, D., Mamyrbayev, O., Zhumagulova, S., Zhumazhan, N. (2024) A Comparative Analysis of LSTM and BERT Models for Named Entity Recognition in Kazakh Language: A Multi-classification Approach. In: Agarwal, N., Sakalauskas, L., Tukeyev, U. (eds) Modeling and Simulation of Social-Behavioral Phenomena in Creative Societies. MSBC 2024. Communications in Computer and Information Science, vol 2211. Springer, Cham. https://doi.org/10.1007/978-3-031-72260-8_10 (in Eng.)

Qin L., Chen Q., Zhou Y., Chen Z., Li Y., Liao L., Li M., Che W., Yu P. S. (2025) A survey of multilingual large language models. Patterns. — Vol. 6, Issue 1, 2025, 101118, ISSN 2666-3899, https://doi.org/10.1016/j.patter.2024.101118 (in Eng.)

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P.J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research. — Vol. 21, issue 1. — P. 5485-5551 (in Eng.)

Rahali A., Akhloufi M.A. (2023) End-to-End Transformer-Based Models in Textual-Based NLP. AI 2023, 4. — P. 54-110. https://doi.org/10.3390/ai4010004 (in Eng.)

Rani Narejo K., Zan H., Oralbekova D., Parkash Dharmani K., Orken M., Mukhsina K. (2024) "Enhancing Emoji-Based Sentiment Classification in Urdu Tweets: Fusion Strategies With Multilingual BERT and Emoji Embeddings," in IEEE Access, vol. 12. — P. 126587-126600, 2024, doi: 10.1109/ACCESS.2024.3446897 (in Eng.)

Rao A., Aithal S., Singh S. (2025) An Evaluation Metric for Assessing Summary-Level Semantic Similarity in Abstractive Text Summarization. 602-607. 10.1109/AIDE64228.2025.10987460 (in Eng.)

Raza A., Soomro M. H., Salahuddin, Shahzad I., Batool S. (2024) Abstractive Text Summarization for Urdu Language. Journal of Computing & Biomedical Informatics, 7(02). Retrieved from https://jcbi.org/index.php/Main/article/view/596 (in Eng.)

Stein R., Jaques P., Valiati J. (2019) An Analysis of Hierarchical Text Classification Using Word Embeddings. Information Sciences. 471. P. 216-232. 10.1016/j.ins.2018.09.001 (in Eng.)

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. (2017) Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010 (in Eng.)

Winarko E., Tanoto L., Reza M.H. (2025) Indonesian Abstractive Text Summarization Using Stacked Embeddings and Transformer Decoder. IAENG International Journal of Computer Science. Vol. 52, Issue 4, April 2025, pp. 1051-1061 (in Eng.)

Zangooei S., Darmani A., Nezhad H., Mahmoudi L. (2025) ARLED: Leveraging LED-based ARMAN Model for Abstractive Summarization of Persian Long Documents. 10.48550/arXiv.2503.10233 (in Eng.)

Zhabayev T., Tukeyev U. (2021) Development of Technology for Summarization of Kazakh Text. International Journal of Advanced Computer Science and Applications (IJACSA), 12(9), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120914 (in Eng.)

Zulkhazhav A., Kozhirbayev Z., Yessenbayev Z., Sharipbay A. (2019). Kazakh Text Summarization using Fuzzy logic. Computación y Sistemas, 23 (in Eng.)

Publication Ethics and Publication Malpractice
in the journals of the Central Asian Academic Research Center LLP

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the journals of the Central Asian Academic Research Center LLP implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The Central Asian Academic Research Center LLP follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the Central Asian Academic Research Center LLP.

The Editorial Board of the Central Asian Academic Research Center LLP will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Ответственный редактор *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Т. Апендиев*
Верстка на компьютере: *Г.Д. Жадырановой*

Подписано в печать 31.03.2026.
Формат 60x881/8.
20,0 п.л. Заказ 1.

«Central Asian Academic Research Center» LLP
Алматы, Шевченко к-сі, 28