УДК 004.89

# АБСТРАКТИВНАЯ СУММАРИЗАЦИЯ НОВОСТЕЙ ВНЕШНЕЙ ТОРГОВЛИ НА ОСНОВЕ НОВОГО СПЕЦИАЛИЗИРОВАННОГО КОРПУСА ДАННЫХ

**Д. А. Лютова[1] [0009-0008-7049-5957], В. А. Малых[2] [0000-0002-4508-2527]**

*[1]Всероссийская академия внешней торговли, г. Москва, Россия*

*[1, 2]Университет ИТМО, г. Санкт-Петербург, Россия*

*[2]Международный университет информационных технологий, г. Алматы, Казахстан*

[1]lyutovad@gmail.com, [2]valentin.malykh@phystech.edu

***Аннотация***

Представлен TradeNewsSum — корпус для абстрактивной генерации аннотаций к новостям внешней торговли, охватывающий русско- и англоязычные публикации из профильных источников. Все рефераты подготовлены вручную по унифицированным правилам. Проведены эксперименты с дообучением трансформерных и seq2seq-моделей и автоматическую оценку по схеме LLM-as-a-judge. Наилучшие результаты показала LLaMA 3.1 в режиме инструкционного промптинга, продемонстрировав высокие значения по метрикам, включая фактологическую полноту.

***Ключевые слова:*** *абстрактивное реферирование, многоязычный корпус, новости внешней торговли, санкции, торговые режимы, TradeNewsSum, трансформеры, большие языковые модели, LLM-as-a-judge, NER-оценка сущностей.*

## ВВЕДЕНИЕ

Автоматическое реферирование становится важным инструментом для обработки новостного потока, особенно в сфере внешней торговли, где требуются краткие и точные изложения сообщений о санкциях, соглашениях и торговых режимах. При этом существующие корпуса в основном англоязычные,

тематически общие и редко содержат качественные абстрактивные аннотации на русском языке, что ограничивает обучение и валидацию многоязычных моделей. TradeNewsSum восполняет этот разрыв: корпус включает тексты на русском и английском языках, собранные с релевантных площадок, и вручную аннотированные краткие рефераты, пригодные для обучения, сравнения и прикладной оценки генерации.

Наш вклад состоит в создании, аннотировании и подробном описании корпуса TradeNewsSum, демонстрации его практической ценности на серии экспериментов с моделями seq2seq, трансформерными и большими языковыми моделями, в том числе в режиме инструкционного промптинга, и сравнительной оценке моделей как классическими метриками ROUGE, BERTScore, NER-F1, так и в парадигме LLM-as-a-judge.

Результаты подтверждают эффективность корпуса для обучения и оценки моделей в задачах автоматического реферирования в домене внешнеэкономических новостей.

## ОБЗОР СУЩЕСТВУЮЩИХ РАБОТ

Для обучения и оценки абстрактивного реферирования широко применяются англоязычные корпуса CNN/DailyMail [20], XSum [14], Newsroom [8], NYT [18] и MultiNews [3], но они слабо адаптированы к внешнеэкономической тематике. В CNN/DailyMail аннотации сгенерированы автоматически и ограничены по качеству [21]; XSum дает однофразовые рефераты с частыми искажениями смысла; Newsroom (≈1.3 млн пар) объединяет разнородные источники и страдает позиционным смещением [11]; MultiNews (56 тыс. примеров) охватывает лишь английский язык [6].

Среди многоязычных наборов XL-Sum [10] (44 языка) фактически использует первое предложение статьи, а MassiveSumm [22] (92 языка) построен автоматически, содержит аннотации низкого качества. Русскоязычные Gazeta [9], RIA [6], Lenta.ru [25] в основном содержат заголовки или метатеги, без полноценных абстрактивных рефератов. MLSum [19] дает ручные аннотации на шести языках (включая русский), но без англоязычных параллелей.

Таким образом, датасеты ограничены тематически, лингвистически и структурно; почти нет полноценных многоязычных корпусов с аннотациями

по внешней экономике, рефераты часто фрагментарны или автоматически извлечены, кросс-языковая оценка затруднена. Русский язык в датасетах представлен слабо. Это снижает применимость существующих моделей к анализу внешнеторговой повестки и подчеркивает необходимость моделей, адаптированных к специализированным многоязычным корпусам.

С учетом этих ограничений особенно важны разработка и адаптация моделей под тематически специализированные и многоязычные корпуса. Ранние решения строились на Seq2Seq с вниманием [1, 17], но качество было достаточно невысоким. Улучшения дали Pointer-Generator [21] и обучение с подкреплением [15], уменьшив повторы и оптимизировав метрики напрямую. Переход к моделям transformer [23] и предобученным энкодер-декодерам (BART [12], T5 [16], PEGASUS [26]) обеспечил значительные результаты на CNN/DailyMail и XSum. Современный этап — это большие языковые модели (БЯМ), например GPT-3 [6], GPT-4, Claude, DeepSeek и др., способные решать задачу по инструкции [7]; при наличии пар предпочтений для донастройки используют DPO, что повышает качество суммаризации [24].

Качество обычно оценивают метриками лексико-семантического сходства — ROUGE [22], METEOR [2], BERTScore [27] — и через извлечение сущностей (NER) для фактологической полноты [4]. Набирают популярность методы с участием БЯМ (GPTScore, G-Eval и аналоги) с более высокой корреляцией с оценками людей [5, 13]; также применяют BLEURT, SummaC, FactCC, MAUVE. Большинство подходов разработано для английского языка и требуют адаптации к задачам на других, включая русский.

**КОРПУС TradeNewsSum**

В рамках проведенного исследования создан специализированный корпус TradeNewsSum[1], ориентированный на абстрактивное реферирование внешнеэкономических новостей. Корпус включает 59395 записей за 2020–2025 г. В него отбирались тексты с содержательной информацией о трансграничных экономических взаимодействиях (экспорт-импорт, инвестиции, санкции, логистические инициативы, гуманитарная помощь и др.).

[1] https://huggingface.co/datasets/lyutovad/TradeNewsSum

Каждая публикация снабжена кратким вручную аннотированным абстрактивным резюме и метаданными: языком оригинала, датой, ссылкой на источник и списком упомянутых стран.

Структура корпуса включает следующие поля:

text — исходный текст публикации,

summary_orig_lang — реферат на языке оригинала,

summary_translated — его перевод на второй язык,

orig_lang — язык оригинала (ru или en),

locations — список стран,

url — источник,

dates — дата публикации.

Корпус является двуязычным: 67% записей представлены на русском языке, 33% — на английском. Для обучения моделей и оценки производительности использовалось стандартное стратифицированное разбиение: обучающая, валидационная и тестовая выборки (80/10/10).

Особенность корпуса — высокая доля абстрактивных аннотаций, сформулированных вручную, что отличает его от большинства русскоязычных ресурсов, основанных на автоматических заголовках.

Сбор новостей проводился из 257 специализированных источников: государственных, агентских, отраслевых и деловых — по темам внешней торговли, санкций, инвестиций и макроэкономики. Основу корпуса составляют публикации на русском и английском языках с официальных сайтов, международных агентств (например, Reuters, Xinhua), деловых СМИ (РБК, «Коммерсантъ»), отраслевых платформ и агрегаторов (UN Comtrade, Eurasianet); актуальность и доступность источников регулярно проверялись вручную.

Тексты варьируются по сложности: от простых однотипных событий в одной стране до многокомпонентных материалов с несколькими странами, товарами и числовыми показателями. Соответственно, аннотации колеблются от кратких до развернутых (до 500–600 знаков), что позволяет обучать модели на разной степени контекстной насыщенности.

Каждая публикация сопровождается составленным специалистом рефератом по следующим формализованным правилам: информационно-деловой

стиль, акцент на сущностях (страны, товары, числовые значения), исключение вводных слов, цитат, оценочных суждений и ссылок; относительные формулировки времени заменяются точными датами по дате публикации. Аннотирование проходило в три этапа: первичная разметка, кросс-проверка и финальное утверждение с участием эксперта; при расхождениях применяется согласование. Все аннотациисоставлены экспертами без автогенерации и извлечения метаданных.

Для корпуса TradeNewsSum рассчитаны количественные характеристики по языкам (русский, английский) и сплитам (обучающая, валидационная, тестовая). В табл. 1 приведены длины текстов и рефератов (в словах и предложениях), показатели словарного разнообразия и лексическое перекрытие между текстами и рефератами.

Средняя длина англоязычных текстов существенно превышает русскоязычные: ≈360 слов против ≈175, при этом средняя длина рефератов остается стабильной — ≈53 слова для английского языка и ≈39 для русского. Количество предложений согласуется с длинами: англоязычные публикации содержат около 17 предложений, русскоязычные — около 11, тогда как рефераты на обоих языках состоят в среднем из 2.6–2.7 предложений. Коэффициент сжатия подтверждает различия: для английского он ниже (≈0.15) по сравнению с русским (≈0.22), что указывает на более агрессивное сжатие англоязычных изложений. В терминах словарного разнообразия английская часть корпуса богаче: абсолютные и средние показатели уникальных слов и лемм выше. Одновременно доля совпадающих лемм между текстом и рефератом больше в русскоязычной выборке, что свидетельствует о более экстрактивном характере русских аннотаций; англоязычные рефераты чаще используют перефразирование и обобщение, что важно учитывать при выборе и настройке моделей.

Для дополнительной оценки качества эталонных аннотаций была расчитана метрика сохранения сущностей (NER-precision, recall, F1). Результаты показывают высокую точность (≈0.85), при относительно низкой полноте (≈0.3), что отражает характер текстов абстрактов: в рефератах сохраняются ключевые акторы, но опускаются второстепенные детали. F1 этих показателей составляет 0.39–0.45.

Табл. 1. Статистика по текстам и рефератам: слова, предложения, леммы и пересечения

| Метрика | Тренировочная часть | | Валидационная часть | | Тестовая часть | |
|---|---|---|---|---|---|---|
| | рус. | англ. | рус. | англ. | рус. | англ. |
| Число пар | 32041 | 15475 | 4005 | 1934 | 4005 | 1935 |
| Мин./Макс. слов (текст) | 8/3335 | 7/4021 | 20/3101 | 11/2764 | 10/3134 | 9/2706 |
| Мин./Макс. слов (реферат) | 5/376 | 7/389 | 6/292 | 8/271 | 6/211 | 8/260 |
| Ср. слов (текст) | 176.0 | 361.5 | 173.5 | 362.9 | 178.4 | 358.4 |
| Ср. слов (реферат) | 39.0 | 52.8 | 38.4 | 53.9 | 39.6 | 52.9 |
| Ср. предлож. (текст) | 1.2 | 16.9 | 11.1 | 16.9 | 11.3 | 16.6 |
| Ср. предлож. (реферат) | 2.6 | 2.6 | 2.6 | 2.7 | 2.7 | 2.6 |
| Коэф. сжатия[2] | 0.222 | 0.146 | 0.221 | 0.149 | 0.222 | 0.148 |
| УЛ[3] (реферат) | 25760 | 21621 | 9792 | 8291 | 10005 | 8153 |
| Совпадающие УЛ | 25405 | 20898 | 9696 | 8107 | 9897 | 8042 |
| Доля новых лемм[4] | 0.014 | 0.033 | 0.01 | 0.022 | 0.011 | 0.014 |

[2] Коэффициент сжатия = Ср. слов (реферат) / Ср. слов (текст)

[3] УЛ – уникальные леммы

[4] Доля новых лемм = 1 - Совпадающие УЛ / УЛ (реферат)

Таким образом, TradeNewsSum представляет собой лингвистически разнородный корпус с элементами как экстрактивного, так и абстрактивного реферирования, что делает его подходящим для обучения seq2seq-моделей и оценки генерации в многоязычном контексте. Благодаря структуре, языковому охвату и качеству аннотаций он может использоваться в задачах реферирования, NER, графового анализа, мониторинга внешней торговли, а также в научных и образовательных целях.

**ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ**

Для оценки качества была использована комбинированная метрика, объединяющая показатели лексико-семантического сходства (ROUGE, METEOR, BERTScore) и фактологической полноты (NER):

$$\text{weighted} = \text{base}_{\text{score}} \cdot \text{ner}_{\text{coef}},$$

где

$$\text{base}_{\text{score}} = 0.20 \cdot \text{BERTScore}_{\text{F1}} + 0.15 \cdot \text{METEOR} + 0.05 \cdot \text{ROUGE}_1 + \\ + 0.10 \cdot \text{ROUGE}_2 + 0.15 \cdot \text{ROUGE}_{\text{Lsum}} + 0.30 \cdot \text{NER}_{\text{F1}},$$

$$\text{ner}_{\text{coef}} = \min\big(1.0, \max(0.6, 0.6 + 0.4 \cdot \text{NER}_{\text{R}})\big).$$

Предложенная метрика weighted учитывает и поверхностное, и семантическое совпадения с эталоном, а также полноту передачи ключевых сущностей, критически важную для новостных задач. Сглаженный штраф $\text{ner}_{\text{coef}}$ повышает устойчивость к частичным пропускам и делает оценку более гибкой. Для оценки без дообучения была использована тестовая выборка TradeNewsSum (5940 текстов, RU+EN). Наилучшие результаты дали mbart-large-cc25 (EN) и mbart_ru_sum_gazeta (RU); pegasus-cnn_dailymail также показала хорошие результаты на английских данных, особенно по NER. ruGPT3 продемонстрировала минимальную точность. Полные значения приведены в табл. 2.

Табл. 2. Результаты моделей без дообучения на тестовой части корпуса

| Модель | язык | weighted | $ROUGE_{Lsum}$ | METEOR | $BERTScore_{F1}$ | $NER_{F1}$ |
|---|---|---|---|---|---|---|
| mT5_multilingual_XLSum | ru | 0.2112 | 0.142 | 0.1849 | 0.8732 | 0.237 |
| | en | 0.2643 | 0.2336 | 0.1801 | 0.8913 | 0.338 |
| mbart-large-cc25 | ru | 0.3867 | 0.2949 | 0.4318 | 0.9079 | 0.47 |
| | en | 0.4252 | 0.4086 | 0.4296 | 0.9084 | 0.494 |
| mbart-large-50-many-to-many-mmt | ru | 0.3098 | 0.2537 | 0.3756 | 0.8932 | 0.329 |
| | en | 0.1325 | 0.0685 | 0.0482 | 0.8361 | 0.067 |
| mbart_ru_sum_gazeta | ru | 0.4464 | 0.4007 | 0.5332 | 0.9259 | 0.528 |
| pegasus-cnn_dailymail | en | 0.3904 | 0.378 | 0.3524 | 0.9015 | 0.493 |
| rugpt3large_based_on_gpt2 | ru | 0.17 | 0.0706 | 0.139 | 0.8408 | 0.118 |
| | en | 0.132 | 0.0741 | 0.0902 | 0.8217 | 0.045 |

Для повышения качества генерации были дообучены модели pointer_generator, mBART, NLLB, mT5 и LLaMA на 47516 парах новостей и аннотаций из корпуса TradeNewsSum. Для LLaMA3:8B-Instruct использовался режим инструкционного инференса без дообучения. Все модели демонстрируют значительное улучшение качества по сравнению с результатами «из коробки», особенно по метрикам ROUGE и $NER_{F1}$. Наивысших значений достигла LLaMA, подтвердив потенциал инструкционного подхода. Подробные результаты представлены в табл. 3.

Табл. 3. Результаты моделей после дообучения на корпусе TradeNewsSum

| Модель | язык | weighted | $ROUGE_{Lsum}$ | METEOR | $BERTScore_{F1}$ | $NER_{F1}$ |
|---|---|---|---|---|---|---|
| pointer_ generator | ru | 0.2916 | 0.0505 | 0.2659 | 0.8261 | 0.391 |
| | en | 0.2559 | 0.0784 | 0.023 | 0.8211 | 0.396 |
| mbart-large-50-many-to-many-mmt | ru | 0.5991 | 0.5093 | **0.7335** | 0.9533 | 0.707 |
| | en | 0.5502 | 0.5427 | 0.5712 | 0.9344 | 0.643 |
| nllb-200-distilled-600M | ru | 0.5948 | 0.49 | 0.7265 | **0.9528** | 0.704 |
| | en | 0.5225 | 0.5178 | 0.5307 | 0.93 | 0.618 |
| mT5_ multilingual_ XLSum | ru | 0.4776 | 0.4143 | 0.5152 | 0.9451 | 0.65 |
| | en | 0.4539 | 0.4911 | 0.4162 | 0.9292 | 0.574 |
| llama3.1:8b-instruct | ru | **0.6269** | **0.5406** | 0.5718 | 0.9448 | **0.728** |
| | en | **0.6099** | **0.583** | **0.6231** | **0.94** | **0.741** |

Для оценки использовался подход LLM-as-a-judge, при котором другая БЯМ анализирует итоговую аннотацию на основе оригинального текста, правил генерации и заданной инструкции. Модель возвращала оценку по критериям точности (faithfulness), соблюдения структуры (structure_adherence), качества языка (style_and_grammar), наличия критических ошибок (critical_violations) и текстового комментария (comment). Такая схема позволяет формализованно оценить качество генерации и выявить потенциальные ошибки без участия человека.

Для снижения вычислительных затрат мы оценивали «медианные» и «сложные» случаи (длинные тексты, высокая числовая насыщенность, множество локаций). Для автоматической оценки итоговых аннотаций использовались две модели: GigaChat Lite и DeepSeek. Модель GigaChat обработала 89.5% примеров (10.5% отказов, главным образом из-за политически чувствительных сюжетов); средние оценки по точности, структуре и стилю превышали 4 балла, критические нарушения не выявлены. Модель DeepSeek оценила все

237 аннотаций, зафиксировав критические ошибки в 36% случаев — в основном из-за неполноты содержания и нарушений структуры (особенно в материалах о санкциях и конфликтах), при этом качество языка стабильно высокое (см. табл. 4).

Табл. 4. Средние оценки качества рефератов по результатам оценки GigaChat и DeepSeek

| Критерий | GigaChat Lite | DeepSeek |
|---|---|---|
| Валидные ответы | 212 (89.5%) | 237 (100%) |
| Отказы от оценки | 10.5% | 0% |
| Точность передачи содержания | 4.11 | 3.69 |
| Структура и следование правилам | 4.11 | 3.85 |
| Язык и стиль | 4.12 | 4.82 |
| Критические нарушения | 0% | 36% (85 из 237) |

Русскоязычные аннотации получают более высокие оценки по точности и структуре, а доля критических нарушений у них значительно ниже: 14.4% против 50.7% для англоязычных текстов. Это указывает на языковую асимметрию качества и подтверждает необходимость дополнительной постобработки англоязычных саммари (табл. 5).

Табл. 5. Сравнение качества рефератов на русском и английском языках

| Показатель | Русский язык | Английский язык |
|---|---|---|
| Точность передачи содержания | 4.23 | 3.32 |
| Структура и следование правилам | 4.11 | 3.66 |
| Язык и стиль | 4.89 | 4.82 |
| Критические нарушения | 14.4% | 50.7% |

## ОГРАНИЧЕНИЯ И НАПРАВЛЕНИЯ РАЗВИТИЯ

Несмотря на высокую проработку, корпус TradeNewsSum имеет ряд ограничений по языковому охвату, тематике и структуре аннотаций. В текущей версии преобладают русско- и англоязычные материалы; доля французских и португальских текстов составляет менее 0.5% и не анализируется в основной части, что сужает возможности многоязычных исследований. Не все сущности (например, компании и товарные группы) отмечены явно, а аннотации не всегда охватывают все фактологические элементы. Отсутствует система версионирования правок, а стратификация по языку усложняет событийный анализ.

В дальнейшем планируется увеличить долю французских и португальских материалов и добавить испанские и китайские новости, ввести явную разметку ключевых сущностей, улучшить методологию аннотирования и внедрить контроль версий, что позволит сформировать полноформатный мультиязычный ресурс.

## ЗАКЛЮЧЕНИЕ

Представлен специализированный корпус TradeNewsSum, ориентированный на генерацию аннотаций к новостям внешней торговли. Корпус охватывает русско- и англоязычные публикации из профильных источников и снабжен подготовленными экспертами рефератами по унифицированным правилам. Проведены эксперименты с дообучением трансформерных и seq2seq-моделей и автоматическая оценка качества по схеме LLM-as-a-judge. Наилучшие результаты продемонстрировала модель LLaMA 3.1 в режиме инструкционного промптинга, показав высокие значения по всем метрикам, включая фактологическую полноту.

Полученные результаты подтверждают применимость предложенного подхода к генерации кратких содержаний в профессиональной новостной повестке. Корпус может использоваться для задач построения дайджестов, анализа торговых потоков и графового моделирования международных отношений. Таким образом, TradeNewsSum является практико-ориентированным ресурсом для исследовательских и аналитических задач, а выявленные ограничения формируют направления его дальнейшего развития.

**СПИСОК ЛИТЕРАТУРЫ**

1. *Bahdanau D. et al*. End-to-end attention-based large vocabulary speech recognition // 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016. P. 4945–4949.

2. *Banerjee S., Lavie A.* METEOR: An automatic metric for MT evaluation with improved correlation with human judgments // Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005. P. 65–72.

3. *Fabbri A. R. et al.* Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model // arXiv preprint arXiv:1906.01749. 2019.

4. *Fischer T., Remus S., Biemann C.* Measuring faithfulness of abstractive summaries // Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022). 2022. P. 63–73.

5. *Fu J. et al.* Gptscore: Evaluate as you desire // arXiv preprint arXiv:2302.04166. 2023.

6. *Gavrilov D., Kalaidin P., Malykh V.* Self-attentive model for headline generation // Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41. Springer International Publishing, 2019. P. 87–93.

7. *Goyal T., Li J. J., Durrett G.* News summarization and evaluation in the era of gpt-3 // arXiv preprint arXiv:2209.12356. 2022.

8. *Grusky M., Naaman M., Artzi Y.* Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies // arXiv preprint arXiv:1804.11283. 2018.

9. *Gusev I.* Dataset for automatic summarization of Russian news // Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9. Springer International Publishing, 2020. P. 122–134.

10. *Hasan T. et al.* XL-sum: Large-scale multilingual abstractive summarization for 44 languages // arXiv preprint arXiv:2106.13822. 2021.

11. *Kryściński W. et al.* Neural text summarization: A critical evaluation // arXiv preprint arXiv:1908.08960. 2019.

12. *Lewis M. et al.* Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension // arXiv preprint arXiv:1910.13461. 2019.

13. *Liu Y. et al.* G-eval: NLG evaluation using gpt-4 with better human alignment // arXiv preprint arXiv:2303.16634. 2023.

14. *Narayan S., Cohen S. B., Lapata M.* Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization // arXiv preprint arXiv:1808.08745. 2018.

15. *Paulus R., Xiong C., Socher R.* A deep reinforced model for abstractive summarization // arXiv preprint arXiv:1705.04304. 2017.

16. *Raffel C. et al.* Exploring the limits of transfer learning with a unified text-to-text transformer // Journal of machine learning research. 2020. Vol. 21, No. 140. P. 1–67.

17. *Rush A.M., Chopra S., Weston J.* A neural attention model for abstractive sentence summarization // arXiv preprint arXiv:1509.00685. 2015.

18. *Sandhaus E.* The New York Times Annotated Corpus Overview [Electronic resource]. Philadelphia: Linguistic Data Consortium, 2008. (LDC Catalog No. LDC2008T19). https://gwern.net/doc/ai/dataset/2008-sandhaus.pdf (accessed: 21.05.2025).

19. *Scialom T. et al.* MLSUM: The multilingual summarization corpus // arXiv preprint arXiv:2004.14900. 2020.

20. *See A., Liu P. J., Manning C.D.* A Neural Attention Model for Abstractive Sentence Summarization [Electronic resource]. 2016. https://github.com/abisee/cnn-dailymail (accessed 07.04.2025).

21. *See A., Liu P.J., Manning C.D.* Get to the point: Summarization with pointer-generator networks // arXiv preprint arXiv:1704.04368. 2017.

22. *Varab D., Schluter N.* MassiveSumm: a very large-scale, very multilingual, news summarisation dataset // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. P. 10150–10161.

23. *Vaswani A. et al.* Attention is all you need // Advances in neural information processing systems. 2017. Vol. 30.

24. *Xin L., Liutova D., Malykh V.* Cross-Language Summarization in Russian and Chinese Using the Reinforcement Learning // International Conference on Analysis of Images, Social Networks and Texts. Cham: Springer Nature Switzerland, 2024. P. 179–192.

25. *Yutkin M.* Lenta.Ru News Dataset [Electronic resource]. 2018. Available at: https://github.com/yutkin/Lenta.Ru-News-Dataset (accessed 04.05.2025).

26. *Zhang J. et al.* Pegasus: Pre-training with extracted gap-sentences for abstractive summarization // International conference on machine learning. PMLR, 2020. P. 11328–11339.

27. *Zhang T. et al.* Bertscore: Evaluating text generation with bert // arXiv preprint arXiv:1904.09675. 2019.

## ABSTRACTIVE SUMMARIZATION FOR TRADE NEWS ANALYSIS BASED ON A NEW DOMAIN-SPECIFIC DATASET

**D. A. Liutova[1] [0009-0008-7049-5957], V. A. Malykh[2] [0000-0002-4508-2527]**

*[1]Russian Foreign Trade Academy, Moscow, Russia*
*[1, 2]ITMO University, Saint Petersburg, Russia*
*[2]International IT University, Almaty, Kazakhstan*
[1]lyutovad@gmail.com, [2]valentin.malykh@phystech.edu

***Abstract***

We present TradeNewsSum—a corpus for abstractive summarization of international trade news—covering Russian- and English-language publications from domain-specific sources. All summaries are manually prepared following unified guidelines. We conducted experiments with fine-tuning transformer and seq2seq models and performed automatic evaluation using the LLM-as-a-judge scheme. LLaMA 3.1 in instruction-prompting mode achieved the best results, showing high scores across metrics, including factual completeness.

**Keywords:** *abstractive summarization, multilingual corpus, international trade news, sanctions, trade regimes, TradeNewsSum, transformers, large language models, LLM-as-a-judge, NER-based entity evaluation.*

**REFERENCES**

1. *Bahdanau D. et al*. End-to-end attention-based large vocabulary speech recognition // 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016. P. 4945–4949.

2. *Banerjee S., Lavie A.* METEOR: An automatic metric for MT evaluation with improved correlation with human judgments // Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005. P. 65–72.

3. *Fabbri A. R. et al.* Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model // arXiv preprint arXiv:1906.01749. 2019.

4. *Fischer T., Remus S., Biemann C.* Measuring faithfulness of abstractive summaries // Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022). 2022. P. 63–73.

5. *Fu J. et al.* Gptscore: Evaluate as you desire // arXiv preprint arXiv:2302.04166. 2023.

6. *Gavrilov D., Kalaidin P., Malykh V.* Self-attentive model for headline generation // Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41. Springer International Publishing, 2019. P. 87–93.

7. *Goyal T., Li J. J., Durrett G.* News summarization and evaluation in the era of gpt-3 // arXiv preprint arXiv:2209.12356. 2022.

8. *Grusky M., Naaman M., Artzi Y.* Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies // arXiv preprint arXiv:1804.11283. 2018.

9. *Gusev I.* Dataset for automatic summarization of Russian news // Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9. Springer International Publishing, 2020. P. 122–134.

10. *Hasan T. et al.* XL-sum: Large-scale multilingual abstractive summarization for 44 languages // arXiv preprint arXiv:2106.13822. 2021.

11. *Kryściński W. et al.* Neural text summarization: A critical evaluation // arXiv preprint arXiv:1908.08960. 2019.

12. *Lewis M. et al.* Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension // arXiv preprint arXiv:1910.13461. 2019.

13. *Liu Y. et al.* G-eval: NLG evaluation using gpt-4 with better human alignment // arXiv preprint arXiv:2303.16634. 2023.

14. *Narayan S., Cohen S. B., Lapata M.* Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization // arXiv preprint arXiv:1808.08745. 2018.

15. *Paulus R., Xiong C., Socher R.* A deep reinforced model for abstractive summarization // arXiv preprint arXiv:1705.04304. 2017.

16. *Raffel C. et al.* Exploring the limits of transfer learning with a unified text-to-text transformer // Journal of machine learning research. 2020. Vol. 21, No. 140. P. 1–67.

17. *Rush A.M., Chopra S., Weston J.* A neural attention model for abstractive sentence summarization // arXiv preprint arXiv:1509.00685. 2015.

18. *Sandhaus E.* The New York Times Annotated Corpus Overview [Electronic resource]. Philadelphia: Linguistic Data Consortium, 2008. (LDC Catalog No. LDC2008T19). https://gwern.net/doc/ai/dataset/2008-sandhaus.pdf (accessed: 21.05.2025).

19. *Scialom T. et al.* MLSUM: The multilingual summarization corpus // arXiv preprint arXiv:2004.14900. 2020.

20. *See A., Liu P. J., Manning C.D.* A Neural Attention Model for Abstractive Sentence Summarization [Electronic resource]. 2016. https://github.com/abisee/cnn-dailymail (accessed 07.04.2025).

21. *See A., Liu P.J., Manning C.D.* Get to the point: Summarization with pointer-generator networks // arXiv preprint arXiv:1704.04368. 2017.

22. *Varab D., Schluter N.* MassiveSumm: a very large-scale, very multilingual, news summarisation dataset // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. P. 10150–10161.

23. *Vaswani A. et al.* Attention is all you need // Advances in neural information processing systems. 2017. Vol. 30.

24. *Xin L., Liutova D., Malykh V.* Cross-Language Summarization in Russian and Chinese Using the Reinforcement Learning // International Conference on Analysis of Images, Social Networks and Texts. Cham: Springer Nature Switzerland, 2024. P. 179–192.

25. *Yutkin M.* Lenta.Ru News Dataset [Electronic resource]. 2018. Available at: https://github.com/yutkin/Lenta.Ru-News-Dataset (accessed 04.05.2025).

26. *Zhang J. et al.* Pegasus: Pre-training with extracted gap-sentences for abstractive summarization // International conference on machine learning. PMLR, 2020. P. 11328–11339.

27. *Zhang T. et al.* Bertscore: Evaluating text generation with bert // arXiv preprint arXiv:1904.09675. 2019.

## СВЕДЕНИЯ ОБ АВТОРАХ

***ЛЮТОВА Дарья Андреевна*** — выпускница магистратуры Университета ИТМО 2025 года по направлению «Искусственный интеллект», аспирантка ИТМО. Исследователь в области обработки естественного языка и аналитики новостных потоков. Область научных интересов: большие языковые модели и методы обработки текста.

***Daria LYUTOVA*** — M.Sc. graduate (2025) in Artificial Intelligence from ITMO University and a Ph.D. student at ITMO. She is a researcher in natural language processing and news analytics. Research interests include large language models and natural language processing.

email: lyutovad@gmail.com
ORCID: 0009-0008-7049-5957

***Малых Валентин Андреевич***, закончил МФТИ в 2009 году. В 2019 году защитил кандидатскую диссертацию по специальности 05.13.11. В настоящее время является доцентом ВШЦК ИТМО, а также профессором-исследователем в МУИТ. Область научных интересов: обработка текстов, большие языковые модели.

***Valentin Malykh*** graduated from MIPT in 2009. In 2019, he defended his PhD in technical sciences. Valentin is currently an assistant professor at the Digital Culture department, ITMO University and research professor at IITU University. Research interests: natural language processing, large language models.

email: valentin.malykh@phystech.edu
ORCID: 0000-0002-4508-2527

*Материал поступил в редакцию 11 октября 2025 года*