# Автоматическое реферирование текстов

В. С. Головизнина
*Вятский государственный университет*
Киров, Россия
golovizninavs@gmail.com

***Аннотация*—Автоматическое реферирование текстов – процесс создания краткого изложения текста, содержащего наиболее важную информацию. В настоящей работе исследуется задача создания рефератов русскоязычных текстов с помощью экстрактивных и абстрактивных методов. Для экспериментов был использован корпус новостных статей Gazeta. Для оценки качества реферирования использовались метрики ROUGE-N, ROUGE-L и BLEU. Наилучшие результаты показала модель ruT5-large.**

***Ключевые слова— автоматическое реферирование текстов, экстрактивные и абстрактивные методы реферирования, mBART, ruGPT-3, ruT5.***

## 1. Введение

Автоматическое реферирование текстов – процесс создания краткого изложения текста, содержащего наиболее важную информацию [1]. Выделяют следующие подходы для реферирования текстов – экстрактивный, абстрактивный и гибридный. При экстрактивном подходе реферат формируется из наиболее важных предложений исходного текста; при абстрактивном содержание реферата генерируется и отличается от предложений исходного текста. Гибридный подход объединяет оба указанных подхода. Методы автоматического реферирования текстов используются в поисковых системах, для резюмирования блогов, научных статей и электронной почты, для генерации заголовков новостных статей, сжатого изложения судебных исков и медицинских текстов [2]. В области автоматического реферирования остаются нерешенные проблемы [2]. Так, в большинстве работ применяется экстрактивный подход, в то время как абстрактивные методы позволяют получить более краткое и близкое к человеческому изложение, отличное от предложений исходного текста. Как и в других областях обработки естественного языка большинство исследований проводится для английского языка [3]. В данной работе сравниваются несколько методов в рамках абстрактивного и экстрактивного подходов на русскоязычном корпусе новостных статей Gazeta.

## 2. Предыдущие работы

Языковые модели, основанные на архитектуре Transformer [4], стали ключевой технологией для решения задач обработки естественного языка, в том числе и для автоматического реферирования текстов [5]. Для реферирования русскоязычных текстов используют модели mBART [6], ruGPT3 [7] и ruT5 [8]. Так, в работе [9] многоязычная модель mBART обучалась задаче реферирования текстов на русскоязычном наборе данных Gazeta. В работе [10] обученная на корпусе Gazeta модель ruGPT-3Small используется для реферирования текстов на русском языке. В работе [11] модель mT5 [12] обучалась реферированию текстов на 44 языках, в том числе и на русском, с использованием корпуса XLSUM.

В данной работе в отличие от [10] вместо модели ruGPT-3Small используется модель ruGPT-3Large. В отличие от [9] помимо mBART обучаются модели ruGPT-3Large и ruT5-large. В отличие от [11] используется не многоязычная модель mT5, а русскоязычная ruT5-large.

## 3. Методы исследования

### *А. Экстрактивные методы*

Для экстрактивного реферирования использовались методы TextRank из библиотеки summa и LexRank из библиотеки lexrank с параметрами по умолчанию.

TextRank [13] – метод, основанный на алгоритме PageRank [14] и применяемый для извлечения ключевых слов и экстрактивного реферирования. LexRank [15] – метод, представляющий текст в виде графа, для вычисления важности текстовых единиц.

### *Б. Абстрактивные методы*

Для абстрактивного реферирования применялись модели mBART, ruGPT-3Large и ruT5-large.

Модель BART – Bidirectional and Auto-Regressive Transformer – основана на архитектуре Transformer и включает в себя двунаправленный кодировщик (как BERT) и авторегрессионный декодировщик (как GPT) [16]. Доступно две версии модели: $BART_{BASE}$ и $BART_{LARGE}$. Многоязычная версия mBART [6] обучалась на Common Crawl corpus для 25 языков. В работе [9] используется mBART, настроенный для реферирования текстов на наборе данных Gazeta.

Модель GPT – Generative Pre-trained Transformer – это 12 слоев декодировщика Transformer [17]. Позднее появились вторая GPT-2 [18] и третья GPT-3 [19] версии модели. Модель ruGPT3 – русскоязычная модель от Сбера [7], основанная на GPT2, доступная в пяти версиях. В работе [20] использовалась модель GPT-2 для реферирования статей о COVID-19 на английском языке. В работе [10] применяли ruGPT-3Small для реферирования статей набора данных Gazeta.

Модель T5 – Text-to-Text Transfer Transformer – обучалась на 24 задачах для английского языка [21]. Модель mT5 обучалась для 101 языка, но на одной задаче – заполнения пропусков в тексте [12]. В работе [11] модель mT5 используется для реферирования на корпусе XLSUM. Модель ruT5 – русскоязычная модель T5 от Сбера, доступная в двух версиях: ruT5-base и ruT5-large [8].

## 4. Результаты исследования

Для моделей $mBART_{LARGE}$ и ruT5-large длина входа (текста) составляла 1024 токена, длина выходных данных (длина генерируемого реферата) ограничивалась максимальной длинной эталонных рефератов.

Для модели ruGPT-3Large длина выходных данных регулировалась так же, длина входных данных – 2048 токенов. При обучении на вход модели ruGPT-3Large подавались последовательности вида: «Text:text[SEP]Summary:summary», где text – это текст, summary – эталонный реферат для этого текста. При тестировании модель генерировала реферат для текста, подаваемого на вход в виде: «Text:text[SEP]Summary:».

Модель $\text{mBART}_{\text{LARGE}}$ дообучена на наборе данных Gazeta в работе [9]. Наборы данных для автоматического реферирования – это наборы текстов и рефератов к ним. Набор данных Gazeta – это 63 435 статей из новостного источника Gazeta.ru. Характеристика набора данных приведена в таблице I (длина в токенах указана для токенизатора ruGPT-3Large). В качестве реферата используется описание статьи. Модели ruT5-large и ruGPT-3Large были дообучены на этом же наборе данных.

Для оценки результатов использовались автоматические метрики: ROUGE-N [22], ROUGE-L [22] и BLEU [23]. Результаты экспериментов приведены в таблице II.

Таблица I. Характеристика набора данных Gazeta

| **Выборка** | **Размер** | **Данные** | **Размер данных в токенах** | | |
|---|---|---|---|---|---|
| | | | ***min*** | ***max*** | ***mean*** |
| Обучающая | 52 400 (82,6%) | текст | 48 | 2 244 | 955 |
| | | реферат | 17 | 123 | 64 |
| Валидационная | 5 265 (8,3%) | текст | 244 | 1 997 | 941 |
| | | реферат | 18 | 124 | 69 |
| Тестовая | 5 770 (9,1%) | текст | 447 | 2 041 | 916 |
| | | реферат | 25 | 67 | 127 |

Таблица II. Результаты экспериментов

| **Метод** | **Значения $F_1$-score** | | | |
|---|---|---|---|---|
| | ***R-1*** | ***R-2*** | ***R-L*** | ***BLEU*** |
| $\text{ruT5}_{\text{LARGE}}$ | **32,45** | **13,97** | **29,24** | 10,88 |
| $\text{ruGPT3}_{\text{LARGE}}$ | 23,45 | 6,45 | 20,73 | 4,93 |
| mBART | 31,55 | 13,54 | 28,22 | **11,19** |
| TextRank | 21,44 | 6,27 | 18,56 | 3,92 |
| LexRank | 23,93 | 8,00 | 20,96 | 5,64 |

## 5. Заключение

Экстрактивный метод LexRank показал более высокие результаты, чем TextRank. Также стоит отметить, что значения метрик метода LexRank выше, чем значения для модели ruGPT-3Large. В рефератах ruGPT-3Large встречаются фактические ошибки, чего не может быть в экстрактивном методе. Лучшие результаты показали модели ruT5-large и mBART. При этом mBART склонен повторять исходный текст. Рефераты ruT5-large краткие и связные, создают наиболее высокое общее впечатление.

## Благодарности

Результаты статьи получены при поддержке грантовой программы Yandex.Cloud.

## Литература

[1] Fabbri, A.R. SummEval: Re-evaluating Summarization Evaluation / A.R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, D. Radev // Transactions of the Association for Computational Linguistics. – 2021. – Vol. 9. – P. 391-409.

[2] EI-Kassas, W.S. Automatic text summarization: A comprehensive survey / W.S. EI-Kassas, C.R. Salama, A.A. Rafea // Expert Systems with Applications. – 2021. – Vol. 165.

[3] Scialom, T. MLSUM: The Multilingual Summarization Corpus / T. Scialom, P. Dray, S. Lamprier, B. Piwowarski, J. Staiano // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2020. – P. 8051-8067.

[4] Vaswani, A. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998-6008.

[5] Liu, Y. Text Summarization with Pretrained Encoders / Y. Liu, M. Lapata // Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. – 2019. – P. 3730-3740.

[6] Liu, Y. Multilingual Denoising Pre-training for Neural Machine Translation / Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer // Transactions of the Association for Computational Linguistics. – 2020. – Vol. 8. – P. 726-742.

[7] Sberbank-ai/ruGPTs [Electronic resource]. — Access mode: https://github.com/sberbank-ai/ru-gpts (01.02.2022).

[8] Sberbank-ai/model-zoo [Electronic resource]. — Access mode: https://github.com/sberbank-ai/ru-gpts (01.02.2022).

[9] Gusev, I. Dataset for Automatic Summarization of Russian / I. Gusev // Artificial Intelligence and Natural Language. – 2020. – P. 122-134.

[10] Nikolich, A. Fine-tuning GPT-3 for Russian Text Summarization / A. Nikolich, I. Osliakova, T. Kudinova, I. Kappusheva, A. Puchkova // Data Science and Intelligent Systems, Proceedings of 5th Computational Methods in Systems and Software. – 2021. – Vol. 2. – P. 748-757.

[11] Hasan, T. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages / T. Hasan, A. Bhattacharjee, M. Islam, K. Samin, Y. Li, Y. Kang, M.S. Rahman, R. Shahriyar // Findings of the Association for Computational Linguistics: ACL-IJCNLP. – 2021. – P. 4693-4703.

[12] Xue, L. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer / L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel // Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2021. – P. 483-498.

[13] Mihalcea, R. TextRank: Bringing Order into Text / R. Mihalcea, P. Tarau // Proceedings of the Conference on Empirical Methods in Natural Language Processing. – 2004. – P. 404-411.

[14] Page, L. The PageRank Citation Ranking: Bringing Order to the Web / L. Page, S. Brin, R. Motwani, T. Winograd. – 1998.

[15] Erkan, G. Graph-based Lexical Centrality as Salience in Text Summarization / G. Erkan, D.R. Radev // Journal of Artificial Intelligence. – 2004. – Vol. 22. – P. 457-479.

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

[17] Radford, A. Improving Language Understanding by Generative Pre-Training / A. Radford, K. Narasimhan, T. Saliman, I. Sutskever. – 2018.

[18] Radford, A. Language Models are Unsupervised Multitask Learners / A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. – 2019.

[19] Brown, T. Language Models are Few-Shot Learners / T. Brown, A. Radford, I. Sutskever. – 2020.

[20] Kieuvongngam, V. Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2 / V. Kieuvongngam, B. Tan, Y. Niu. – 2020.

[21] Raffel, C. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer / C. Raffel, N, Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu // Journal of Machine Learning Research. – 2020. – Vol. 21. – P. 1-67.

[22] Lin, C. ROUGE: A Package for Automatic Evaluation of Summaries / C. Lin // Association for Computational Linguistics. – 2004. – P. 74-81.

[23] Papineni, K. BLEU: a Method for Automatic Evaluation of Machine Translation / K. Papineni, S. Roukos, T. Ward, W. Zhu // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. – 2002. – P. 311-318.