УДК: 004.912.4
ГРНТИ: 28.23.25

Молчанов А.Д.
Магистр
2 курс, направление «Информатика и вычислительная техника»
Национальный исследовательский технологический университет «МИСиС»
Россия, г. Москва

ОБЗОР МЕТОДОВ АВТОМАТИЧЕСКОГО УПРОЩЕНИЯ ТЕКСТА

Статья посвящена обзору методов и подходов к автоматическому упрощению текста (УТ, Text Simplification). Основное применение данные методы находят в снижении лингвистической сложности текста при сохранении его смыслового содержания, что, в свою очередь, находит востребование среди людей, изучающих иностранный язык, а также людей с дислексией, афазией или расстройствами аутистического спектра. Если ранние подходы опирались на ручные правила и извлечение предложений, то современные методы с большей частотой используют методы глубокого обучения. В данной обзорной статье рассматриваются эволюция подходов к УТ и ключевые подходы (лексическое, синтаксическое упрощение, нейросетевая генерация).

Ключевые слова: Упрощение текста, лексическое упрощение, синтаксическое упрощение, глубокое обучение, seq2seq-модели, SARI, BLEU.

The article is devoted to an overview of methods and approaches to automatic text Simplification. The main application of these methods is to reduce the linguistic complexity of the text while preserving its semantic content, which, in turn, is in demand among people learning a foreign language, as well as people with dyslexia, aphasia or autism spectrum disorders. While early approaches relied on manual rules and sentence extraction, modern methods use deep learning methods with greater

frequency. This review article examines the evolution of approaches to UT and key approaches (lexical, syntactic simplification, neural network generation).

Keywords: Text simplification, lexical simplification, syntactic simplification, deep learning, seq2seq models, SARI, BLEU.

Ведение:

В узкой трактовке терминологии, упрощение текста — это процесс снижения лексической сложности с условием сохранения ключевой информации и смысла оригинального текста. В более широкой интерпретации, оно включает концептуальное упрощение, может включать разъяснение и даже элементы пересказа. УТ тесно связано с областями машинным переводом, сокращением и генерацией текста [1], но, в отличие от сокращения, обычно сохраняет весь объём содержания и может включать случаи, когда итоговый текст может быть длиннее исходного.

Слои населения, которые могут извлечь пользу из УТ разнообразны: изучающие иностранный язык [2], которым может потребоваться упрощение текста на лексическом уровне, люди с дислексией [3], у которых трудности вызывают длинные слова и определённые буквосочетания, люди с расстройствами аутистического спектра, которым может требоваться снижение числа фигуральных выражений и синтаксической сложности, а также начинающие читатели. В связи с экспоненциальным ростом объемов неструктурированных текстовых данных, наблюдается значительное повышение исследовательского и практического интереса к технологиям УТ на протяжении последних полутора десятилетий.

Целью представленной обзорной статьи является анализ современных методов упрощения текста, для их потенциального последующего употребления в разработке ПО для такой задачи.

Материалы и методы

На данный момент, существующие методы УТ можно разделить на две категории:

Экстрактивные - более ранние методы, которые выбирают из текста предложения с наибольшей смысловой нагрузкой, например, на основе статистической меры TF-IDF. Фактически, эти методы представляют собой методы сокращения. Их основное достоинство заключается в простоте их реализации, а недостатки не порождают нового текста и не упрощают лексику или синтаксис.

Абстрактные – в отличии от экстрактивных методов генерируют новый, более простой текст. Абстрактные методы могут быть узко специализированными, фокусируясь только на лексических или фразовых заменах для упрощения текста на уровне отдельных предложений [4], пренебрегая задачами грамматического и синтаксического упрощения [5]. Более продвинутые итерации этих методов добавления и удаления текста, а также разбиения предложений для их упрощения. Наиболее распространенным подходом является нейросетевая архитектура последовательность в последовательность, RNN и LSTM. Такие методы могут быть подразделены на две подкатегории в зависимости от их задачи:

Лексическое упрощение (ЛУ) - замена сложных слов и фраз более простыми альтернативами.

Генерация нового текста - синтаксическое упрощение, статистический машинный перевод, нейросетевые архитектуры последовательность в последовательность, в том числе с механизмами внимания, указательно-генеративными сетями, обучением с подкреплением и т.д.

В соответствии с целью данной статьи основное внимание уделяется абстрактным методам, особенно нейросетевым, как наиболее перспективным для локального применения. В качестве метрик оценки используются BLEU – метрика для оценки схожести выходных данных автоматического упрощения с эталонными, ручными упрощениями, SARI [6] - специализированная метрика

для УТ, оценивающая добавления, удаления и сохранения слов, а также автоматические индексы удобочитаемости, как например индекс удобочитаемости Флеша-Кинкейда.

Подходы к упрощению текста

Экстрактивный подход

Экстрактивное упрощение представляет собой примененные для упрощения методы сокращения: выбор предложений, передающих наибольший «смысл». Простейший пример - метод TF-IDF с последующим вычислением веса предложения - сумма TF-IDF слов, делённая на длину. Порог или N предложений с наибольшим весом отбираются для получения результирующего текста. Ещё одним недостатком данного подхода является необходимость в предобработке входного текста. Как уже говорилось выше, методы этого подхода просты в реализации, но не генерируют нового текста и не упрощают лексику и синтаксис.

Абстрактный подход - лексическое упрощение

Лексическое упрощение, впервые исследованное и описанное Девлином и Тейтом [7], представляет собой форму абстрактного УТ, нацеленную на замену сложных, для выбранной целевой аудитории, слов на более простые альтернативы. Эффективность лексического упрощения в значительной степени зависит от используемой базы данных и критериев выбора подходящей замены. Алгоритм выполнения лексического упрощения можно разделить на следующие четыре этапа:

Выявление сложных слов - выполняется перед всеми остальными этапами и определяет, какие именно слова требуется упростить для предполагаемой целевой аудитории. Ранние системы УТ ([7]) пропускали данный этап, исходя из утверждения что все слова могут подлежать упрощению. Но такой подход являлся как ресурсо не эффективным, так и терял смысловое наполнение в около половины случаев. Шардлоу в исследовании [5] доказал, что пропуск данного шага делает результирующий текст безграмотным или бессвязным.

Современные стратегии лексического упрощения можно подразделить на следующие категории: пороговые методы (по частоте слов), лексиконные (терминология в целевой области), неявные (отбраковка более сложных замен) [6] и машинное обучение.

Генерация замен - на этом этапе “создаются” замены для всех определённых сложных слов. В идеале, должен находить все возможные варианты упрощения, из которых неподходящие будут отсеяны далее. Более ранние системы извлекали синонимы, гиперонимы и парафразы из лингвистических баз на подобие WordNet, PPDB, Simple PPDB. Современные систему чаще используют извлечение замен из параллельных сложно-простых корпусов как более эффективную и быструю альтернативу. Наиболее популярный ресурс – это Simple English Wikipedia [8], состоящий из пар англоязычных статей с Wikipedia.

Выбор замены – на данном этапе происходит отбрасывание менее подходящих замен по средствам снятие лексической неоднозначности, POS-тегирования и фильтрации по семантической близости.

Ранжирование замен - частотный подход, комбинированные метрики, SVM и нейросетевые решения.

Основными нерешёнными проблемами лексического упрощения являются проблема замены многозначных слов без потери контеста и искажения смысловой нагрузки [5], замена фраз и необходимость в доменно-специализированных ресурсах.

Абстрактный подход - генерация нового текста

Синтаксическое упрощение

В отличие от лексического упрощения, работающего только на лексическом уровне, синтаксическое упрощение выявляет и работает с грамматически сложным текстом. Операции упрощения, выполняемые им, включают, но не ограничиваются: разбиение сложных предложений на несколько простых, перевод из пассивного залога в активный, разрешение анафоры.

Алгоритм его выполнения может быть разбит на три основные фазы: анализ – построение дерева разбора, трансформация - разбиение предложений, перестановка и удаление слов, и “регенерация” – этап служащий для повышения связности и читаемости результирующего текста. Ранние системы использовали ручные правила изменения текста, что, хоть и достигало высокой точности, было очень трудозатратно.

Статистический машинный перевод

В рамках данного подхода, задача УТ представляется как перевод со «сложного» языка на «простой». Для решения этой задачи используются модифицированные стандартные системы статистического машинного перевода, как например Moses, дополненные модулями удаления фраз. Word embeddings и seq2seq модели улучшили качество, но всем моделям данного подхода свойственны невозможность учёта пунктуации и сложности с разбиением длинных предложений, вызванные механизмами самих методов статистического машинного перевода.

Современные методы глубокого обучения

Первыми шагами в направлении современных методов УТ можно считать предложенную Вангом с соавторами [9] LSTM модель для замены, удаления и перестановки слов. Однако у этой и других ранних моделей последовательность-в-последовательность наблюдалось несколько критических недостатков, а именно неточность воспроизведения и повторения. Первое характеризуется сложностями в размещении редких слов в результирующем тексте, что в последствии было решено добавлением указателей во входной текст. Второе – слишком частым предсказанием стоп-слов (слов без смысловой нагрузки - предлогов, союзов, местоимений), приводящим к их повторению и потери смысловой нагрузки, что было решено добавлением штрафа с помощью вектора, отслеживающего внимание и контекст.

Обучение с подкреплением – предложенная Зангом и др. [10] модель, совместно формирующая простоту, грамматичность и семантическую точность,

с дополнительным компонентом в виде лексического упрощения. Этот тип обучения позволяет вводить априорные знания в задачу упрощения.

Memory-augmented NSE. By с соавторами [11] и предложили архитектуру Neural Semantic Encoder с дополненной памятью, значительно снижающую сложность текста при сохранении грамматики и смысловой нагрузки.

Обучаемые без учителя методы [12] используют инициализацию фразовых таблиц и раздельные кодировщики-декодировщики без параллельных данных.

Таблица 1.

Сравнение моделей УТ

Модель	Датасет	BLEU	SARI
SBMT-SARI [6]	TurkCorpus	~37.0	~33.1
Dress-Ls [10]	WikiLarge	~39.0	~37.2
NSELSTM-B [11]	WikiLarge	~33.4	~37.9
Unsupervised NTS [12]	ASSET / Newsela	~34.0	~35.0
BART-base [13]	WikiLarge	~30.9	~38.3

Выводы:

В вышеизложенной обзорной статье была прослежена эволюция методов упрощения текста от ранних экстрактивных подходов, основанных на TF-IDF и ручных правилах, до современных абстрактных методов с использованием глубокого обучения. Было выявлен, что ключевым ограничением остаётся дефицит высококачественных лингвистических баз и параллельных корпусов на подобие Simple English Wikipedia [8], который частично компенсируется созданием синтетических данных и разработкой методов, не требующих параллельной разметки. Сравнительный анализ моделей на основе метрик BLEU и SARI демонстрирует, что наилучшие результаты на сегодняшний день показывают архитектуры семейства BART и модели с обучением с подкреплением (Dress-Ls), тогда как методы обучения без учителя, хотя и

уступают им в точности, предоставляют возможности для упрощения текстов на языках с ограниченными ресурсами.

Использованные источники:

1. Zhemin Zhu et al. A Monolingual Tree-based Translation Model for Sentence Simplification // – 2010. – Режим доступа: https://www.researchgate.net/publication/221102865_A_Monolingual_Tree-based_Translation_Model_for_Sentence_Simplification (дата обращения: 01.04.2026).
2. Jun Liu, Yuji Matsumoto. Simplification of Example Sentences for Learners of Japanese Functional Expressions. // – 2016. – Режим доступа: https://www.aclweb.org/anthology/W16-4901 (дата обращения: 01.04.2026).
3. LuzRello, et al. Frequent words improve readability and short words improve understandability for people with dyslexia. // – 2013. – Режим доступа: https://doi.org/10.1007/978-3-642-40498-6_15 (дата обращения: 01.04.2026).
4. Gustavo H Paetzold et al. A survey on lexical simplification. // Journal of Artificial Intelligence Research 60. – 2017. – С. 549–593. (дата обращения: 03.04.2026).
5. Matthew Shardlow. A Survey of Automated Text Simplification. // – 2014. – Режим доступа: https://doi.org/10.14569/SpecialIssue.2014.040109 (дата обращения: 04.04.2026).
6. Wei Xu et al. Optimizing Statistical Machine Translation for Text Simplification // – 2016. – Режим доступа: https://www.researchgate.net/publication/329975918_Optimizing_Statistical_Machine_Translation_for_Text_Simplification (дата обращения: 03.04.2026).
7. Siobhan Devlin and John Tait. The use of a psycholinguistic database in the simple cation of text for aphasic readers. // -1998. (дата обращения: 06.04.2026)

8. Mark Yatskar et al. For the sake of simplicity: Unsuper vised extraction of lexical simplifications from Wikipedia. // - 2010. arXiv:1008.1986 – Режим доступа: https://arxiv.org/abs/1008.1986 (дата обращения: 09.04.2026)
9. Tong Wang et al. An Experimental Study of LSTM Encoder-Decoder Model for Text Simplification. // - 2016. arXiv:1609.03663 – Режим доступа: http://arxiv.org/abs/1609.03663 (дата обращения: 10.04.2026)
10. Xingxing Zhang et al. Sentence Simplification with Deep Reinforcement Learning. // - 2017. arXiv:1703.10931 – Режим доступа: 1703.10931 (дата обращения: 12.04.2026)
11. Tu Vu et al. Sentence Simplification with Memory-Augmented Neural Networks. // - 2018. arXiv:1804.07445 – Режим доступа: 1804.07445 (дата обращения: 12.04.2026)
12. Sai Surya et al. Unsupervised Neural Text Simplification. // - 2019. arXiv:1810.07931 – Режим доступа: 1810.07931 (дата обращения: 13.04.2026)
13. Tshering et al. Text Simplification Using T5 Model and BART Model. // - 2025 – Режим доступа: https://www.researchgate.net/publication/390854431_Text_Simplification_Using_T5_Model_and_BART_Model (дата обращения: 13.04.2026)