УДК 81'322.2

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

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

ОЦЕНКА КАЧЕСТВА РУССКОЯЗЫЧНЫХ АВТОМАТИЧЕСКИХ РЕФЕРАТОВ С ПРИМЕНЕНИЕМ МЕТРИК НА БАЗЕ N-ГРАММ (на материале научных текстов из области ИКТ)

В данной статье проведен анализ применимости основанных на n-граммах метрик BLEU и ROUGE для оценки автоматически сгенерированных рефератов русскоязычных научных статей в области информационно-коммуникационных технологий (ИКТ). Выделены связанные со спецификой функционирования метрик проблемы, касающиеся русского языка и русскоязычных научных текстов, в частности, текстов из области ИКТ. Сделан вывод о недостаточной достоверности предоставленной данными метриками оценки качества русскоязычных научных рефератов ИКТ-текстов, связанной с лексико-морфологической сложностью русского языка и самой спецификой функционирования автоматических метрик на базе n-грамм.

Ключевые слова: автоматическое реферирование; информационно-коммуникационные технологии; научный дискурс; n-граммы; автоматические метрики; BLEU; ROUGE.

QUALITY EVALUATION OF RUSSIAN -LANGUAGE AUTOMATIC SUMMARIES WITH N-GRAM METRICS (based on ICT Scientific Texts)

This article analyzes the applicability of n-gram-based metrics (BLEU and ROUGE) for evaluating automatically generated summaries of Russian-language scientific papers in the field of Information and Communication Technologies (ICT). The problems associated with the specifics of the metrics' functioning, concerning the Russian language and Russian-language scientific texts, in particular, texts from the field of ICT, are identified. The study concludes that the quality assessment of Russian-language scientific summaries of ICT texts, as provided by these metrics, lacks sufficient reliability. This limitation stems from the lexico-morphological complexity of the Russian language and the inherent constraints of n-gram-based automatic evaluation metrics.

Key words: automatic summarization; information and communication technologies; scientific discourse; n-grams; automatic metrics; BLEU; ROUGE.

Экспоненциальный рост объема научных публикаций обусловливает популярность автоматического реферирования как инструмента обработки научной информации. Особую актуальность оно приобретает в быстроразвивающихся дисциплинарных областях, к числу которых относятся информационно-коммуникационные технологии (ИКТ), где важен оперативный анализ

значительных массивов текстовой информации. В академической среде автоматическое реферирование способно ускорить обработку больших массивов публикаций, помогая исследователям быстрее находить релевантные работы.

В контексте активного внедрения нейросетевых технологий для автоматического реферирования в научно-исследовательскую практику особую актуальность приобретает проблема оценки и контроля качества генерируемых рефератов. Так, эталонный реферат научного текста должен:

1) сохранять содержательную адекватность: точная передача ключевых положений исходного текста при сохранении смысловой целостности и научной достоверности информации обеспечивает корректное восприятие научного контента;

2) соблюдать нормативы академического стиля: сохранение точной терминологии, логической последовательности изложения и соответствие жанровым нормам научной коммуникации.

Рефераты, генерируемые большими языковыми моделями, являются преимущественно абстрактивными. Такой тип реферирования слабо поддается контролю со стороны пользователя. В этой связи особую актуальность приобретает задача оценки качества автоматических абстрактивных рефератов на русском языке, характеризующемся высокой степенью синтетичности морфологии и синтаксиса, а также гибким порядком слов в предложениях.

Для автоматического контроля качества автоматического реферирования исследователями активно применяются метрики, основанные на n-граммах (последовательностях, состоящих из n элементов, которые могут быть звуками, слогами, словами или буквами, в зависимости от контекста), такие как BLEU и ROUGE. Спецификой данного типа метрик является: 1) учет локальных последовательностей (n-граммы анализируют последовательности из n подряд идущих элементов (слов, символов), что позволяет оценивать устойчивые статистические зависимости или повторяющиеся структуры, которые проявляются на ограниченных участках текста, обычно в пределах небольшого числа последовательно идущих элементов (слов, букв)); 2) статистическая природа получаемых результатов.

Эмпирическую базу исследования составили два реферата отобранной из электронной научной библиотеки «КиберЛенинка» статьи, посвященной методам автоматического анализа темпоральных выражений в текстах на естественном языке: 1) созданный вручную эталонный реферат, наличие которого является непременным условием для работы данных метрик и проверки их валидности (далее – эталон); 2) автоматический реферат, сгенерированный большой языковой моделью DeepSeek-V3 (далее – реферат).

Выбор статьи осуществлялся по следующим критериям: 1) соответствие тематической области ИКТ; 2) большой объем (45 341 знак); 3) наличие элементов, типичных для научных текстов в ИКТ-дискурсе: формул, таблиц, алгоритмов, насыщенность аббревиатурами преимущественно на латинице.

Инструментарий исследования представлен автоматическими метриками BLEU и ROUGE.

Целью данного исследования является определение эффективности и ограничений применения основанных на n-граммах метрик для оценки качества автоматически генерируемых рефератов русскоязычных научных текстов в области ИКТ.

Принципиальным **условием** исследования явилось моделирование типичной пользовательской ситуации при автоматическом реферировании: отсутствие предварительной обработки исходного текста, минимально детализированный запрос к модели (без составления сложных промптов), отсутствие постобработки сгенерированного реферата.

В ходе исследования сгенерированный нейросетью реферат был оценен автоматически по метрикам BLEU и ROUGE. Далее полученные по каждой из метрик данные подверглись анализу в контексте того факта, что изначально эти метрики были разработаны для английского языка, имеющего структуру принципиально отличную от русского языка. Необходимость использования данных метрик обусловлена отсутствием лингвистически адаптированных для русского языка метрик оценки качества автоматических рефератов.

1. Метрика BLEU (Bilingual Evaluation Understudy)

Данная автоматическая метрика была изначально разработана специалистами IBM в 2002 г. для оценки качества машинного перевода [1], но сегодня часто находит свое применение и в решении других задач генерации текста, в частности, оценке качества автоматического реферирования. Она сравнивает сгенерированный текст с эталонными референсами (в случае оценки качества автоматического реферирования эталоном будет созданный вручную реферат) на основе n-граммной статистики и оценивает точность совпадения n-грамм между автоматическим рефератом и эталоном. Эта метрика работает с последовательностями слов длиной 1–4 (униграммы (отдельные слова), биграммы (два последовательно идущих слова), триграммы (три последовательно идущих слова), кватрограммы (четыре последовательно идущих слова), чувствительна к пропускам и перестановкам слов, неточным формулировкам, а также штрафует за слишком короткие формулировки.

Данные расчетов по метрике BLEU составили (табл. 1):

Таблица 1

Оценка автоматического реферата по метрике BLEU

Метрика	Оценка
Униграммы	0,75
Биграммы	0,62
Триграммы	0,55
Кватрограммы	0,40
ИТОГО	0,58

Метрика BLEU ранжируется от 0 при полном несовпадении до 1 при абсолютной идентичности с эталоном. Необходимо отметить, что с учетом строгости метрики при оценке абстрактивного реферата показатель 1 по умолчанию недостижим. Исходя из этого, результат 0,58 уже свидетельствует о формально приемлемом качестве сгенерированного реферата.

Важным будет рассмотреть примеры участков текста, получивших по данной метрике высокий, средний и низкий балл.

Пример 1 : Высокий балл (0,7–1,0)

(эталон) *Стандарт ISO-TimeML утвержден ISO в 2009 году*

(реферат) *ISO-TimeML стал международным стандартом в 2009 г.*

Результат:

Униграммы: все слова присутствуют

Биграммы: 3 совпадения (*ISO-TimeML*, *2009 г.*, *международным стандартом*)

Триграммы: 1 совпадение (*в 2009 году* ≈ *в 2009 г.*)

Кватрограммы: 0 совпадений (*стандарт ISO-TimeML утвержден* ≠ *ISO-TimeML стал международным*)

Штраф: небольшой штраф за сокращение *году* → *г.*

Итоговый BLEU: ~0,65.

Интерпретация результата: несмотря на перефразирование, ключевые элементы сохранены.

На уровне униграмм, биграмм и триграмм работа BLEU вполне корректна. Далее при переходе к анализу кватрограмм проявляется особенность метрики, связанная с ее высокой чувствительностью к перефразированию: фактически предложение из автоматического реферата полностью корректно, однако высокий балл оно получает за исключительно техническое совпадение некоторого количества элементов без учета их положения в предложении и смысла, который они несут.

Пример 2 : Средний балл (0,3–0,6)

(эталон) *Система HeidelTime достигает F-меры 86 % на Task A TempEval-2*

(реферат) *Точность HeidelTime составляет 86 %*

Результат:

Униграммы: 2 совпадения (*HeidelTime, 86 %)*

Биграммы: 1 совпадение (*86 %*)

Триграммы: 0 совпадений

Кватрограммы: 0 совпадений

Пропущено: *F-мера, TaskATempEval-2*

Итоговый BLEU: ~0,35.

Интерпретация результата: наблюдается потеря ключевых терминов.

В данном случае применение строгой метрики позволило обеспечить контроль над сохранностью терминологического состава: метрика абсолютно верно зафиксировала некорректную замену *F-меры* на *точность* (F-мера представляет собой среднеравновесное между точностью и полнотой). Однако следует подчеркнуть, что это актуально преимущественно в отношении англоязычных заимствований и аббревиатур, обладающих фиксирован-

ной формой. В отличие от них, русскоязычные термины подчиняются морфологическим законам русского языка, предполагающим их грамматическое изменение в соответствии с синтаксическим и семантическим контекстом конкретного высказывания.

Пример 3 : Низкий балл (0–0,2)

(эталон) *Для русскоязычных текстов отсутствуют размеченные корпуса, аналогичные TimeBank.*

(реферат) *Для русского языка нет обучающих данных.*

Результат:

Униграммы: 1 совпадение (*Для*)

Итоговый BLEU: 0,1.

Интерпретация результата: практически полное несоответствие.

Здесь низкий балл обусловлен совпадением только одной униграммы. Возникает явный диссонанс: формально низкий балл за точность при фактической корректности предложения в сгенерированном реферате, хоть и содержащего обобщение (*размеченные корпуса, аналогичные TimeBank* → *обучающих данных*).

2. Группа метрик ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Данная группа метрик была предложена ученым в области информатики Chin-Yew Lin [2] в 2004 г. с целью приблизить качество автоматической оценки рефератов к качеству ручного оценивания. В отличие от BLEU, ориентированной на точность, ROUGE оценивает полноту (recall) охвата ключевой информации из исходного текста на основе подсчета совпадений n-грамм (последовательностей слов) между рефератом и эталоном.

В настоящем исследовании из группы метрик были использованы ROUGE-1 (совпадение униграмм (отдельных слов)), ROUGE-2 (совпадение биграмм (пар слов), ROUGE-L (совпадение длинных подпоследовательностей слов – непрерывных или разрывных цепочек слов, сохраняющих порядок следования и встречающихся как в автоматическом реферате, так и в эталоне).

Оценочные показатели ROUGE для автоматического реферата представлены следующим образом (табл. 2):

Таблица 2

Оценка автоматического реферата
по автоматическим метрикам ROUGE

Метрика	**Содержание метрики**	**Оценка**
ROUGE-1	совпадение униграмм	0,78
ROUGE-2	совпадение биграмм	0,65
ROUGE-L	совпадение длинных подпоследовательностей слов	0,72

Как и для BLEU, показатели данной группы метрик также стремятся к 1, по умолчанию недостижимой при оценке качества абстрактивного реферата. Согласно группе метрик ROUGE автоматический реферат, имеющий достаточно высокие показатели, является качественным.

Однако при детальном анализе принципа работы метрик обнаруживаются существенные нюансы.

ROUGE-1 технически подсчитывает совпадение униграмм вне зависимости от окружающего их контекста.

Пример 1:
(эталон) *дейктические элементы*
(реферат) *обработка дейктических элементов*
Результат: униграмма *дейктические* засчитана.

Интерпретация результата: высокий показатель сходства (0,78) в данном случае свидетельствует исключительно о частичном совпадении униграмм сгенерированного реферата с эталонным без учета семантической и синтаксической корректности расположения единиц. Очевидно, что система распознает основы слов.

С учетом высокой флективности русского языка, существует значительная вероятность ошибочной положительной идентификации морфологически некорректных словоформ в автоматически сгенерированном реферате, что может привести к искусственному завышению оценки качества текста.

Тексты научного дискурса отличаются повышенной синтаксической и морфологической сложностью, терминология в области ИКТ характеризуется смешением русского и английского языков и высокой степенью аббревиации, что создает дополнительные трудности для метрических систем и повышает вероятность некорректной оценки качества автоматического реферата.

ROUGE-2 технически анализирует биграммы вне контекста.

Пример 2:
(эталон) *интервальная логика Аллена*
(реферат) *логика временных интервалов*
Результат: совпадающих биграмм нет.

Интерпретация результата: 0 баллов, однако фактически данные выражения являются контекстуальными синонимами, сохраняющими идентичный смысл.

Оценка (0,65) показывает наличие в тексте автоматического реферата некоторого количества совпадающих пар слов. При этом такой показатель не гарантирует корректность их употребления в реферате и значимость для понимания смысла исходного текста.

Свойственная русскому языку лексико-синтаксическая вариативность и отсутствие фиксированного порядка слов при изначальной англоязычной ориентированности метрики детерминируют высокую вероятность некор-

ректной идентификации значимой информации, а также учета биграмм, имеющих в составе служебные слова. Это, в свою очередь, может обусловить неверную оценку качества автоматического реферата.

ROUGE-1 + ROUGE-2. В некоторых случаях в работе двух сопряженных метрик наблюдается противоречие при оценке одного и того же участка текста.

Пример 3 (уровень предложения):

(эталон) *Система HeidelTime использует правило-ориентированный подход.*

(реферат) *HeidelTime применяет правила.*

Результат: ROUGE-1 даст высокий балл за *HeidelTime* и *правила*, но ROUGE-2 снизит балл за пропуск биграммы *правило-ориентированный подход*.

ROUGE-L технически оценивает длину наибольшей общей подпоследовательности. Метрика не требует строгого непрерывного совпадения. Это обеспечивает большую устойчивость к небольшим перефразированиям или добавлению служебных слов. ROUGE-L учитывает сохранение последовательности ключевых элементов, что важно для научных текстов, где логика изложения критична.

Пример 4:

(эталонная последовательность) *Описание проблем. – Методы. – Примеры систем*

(последовательность в реферате) *Проблемы. – Методы. – (Примеры опущены)*

Результат: частичное совпадение структуры→балл снижен.

Интерпретация результата: результат 0,72 свидетельствует о том, что технически структура автоматического реферата приблизительно соответствует структуре эталона.

Анализ принципа работы автоматических метрик ROUGE и BLEU позволяет выделить ряд проблем, связанных со спецификой функционирования метрик, основанных на статистическом учете n-грамм.

Проблемы, обусловленные спецификой русского языка

1. Гибкий порядок слов в русском языке снижает точность n-граммных метрик, так как они жестко привязаны к последовательности элементов.

2. Морфологическая вариативность детерминирует занижение оценки BLEU из-за несовпадения словоформ, даже если лексическая база идентична.

3. Контекстуальная синонимия и многозначность терминов не учитываются ни BLEU, ни ROUGE, что приводит к необоснованному занижению оценок при корректном перефразировании.

Проблемы, обусловленные применением метрик для научных текстов

4. В научном тексте допускается вариативность формулировок, но метрики, основанные на n-граммах, снижают оценки за отклонения от эталонного реферата.

5. Оценка исключительно лексического покрытия без учета смысловой связности нецелесообразна для научного реферата, требующего соблюдения

четкой структуры и логики. Также статистические метрики не различают стилевую окраску слов, что критично для оценки качества реферата научной статьи.

Проблемы, обусловленные спецификой текстов в области ИКТ

6. Научные тексты в области ИКТ характеризуются высокой терминологической плотностью, многие составные термины имеют сложную структуру. Термины такого типа BLEU может посчитать ошибочными при перефразировании или изменении морфологической формы.

7. Технические формулы и код: Ни BLEU, ни ROUGE не приспособлены для оценки вставок математических выражений или фрагментов программного кода, которые часто встречаются в ИКТ-рефератах.

Анализ выявленных проблем позволяет сделать **вывод** о том, что метрики, разработанные для английского языка с его фиксированным порядком слов и меньшей морфологической вариативностью, не предоставляют достоверную информацию о качестве русскоязычного научного реферата, включая даже базовые статистические показатели: оценки могут быть завышены либо занижены. Это связано с лексико-морфологической сложностью русского языка и самой спецификой функционирования автоматических метрик на базе n-грамм.

Для адекватной оценки рефератов научных текстов в области ИКТ необходима разработка специализированных метрик, учитывающих как синтетический характер русского языка, так и особенности научного дискурса ИКТ.

ЛИТЕРАТУРА

1. BLEU: a method for automatic evaluation of machine translation / K. Papineni, S. Roukos, T. Ward, W. J. Zhu // ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. 2002. P. 311–318.
2. Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. Barcelona : Association for Computational Linguistics. 2004. P. 74–81.