УДК 004.8.81 doi: 10.15827/0236-235X.142.228-236 2023. Т. 36. № 2. С. 228–236

Использование языковых моделей T5 для задачи упрощения текста

Д.Д. Васильев
А.В. Пятаева

Ссылка для цитирования
Васильев Д.Д., Пятаева А.В. Использование языковых моделей T5 для задачи упрощения текста // Программные продукты и системы. 2023. Т. 36. № 2. С. 228–236. doi: 10.15827/0236-235X.142.228-236
Информация о статье
Поступила в редакцию: 13.03.2023 После доработки: 14.03.2023 Принята к публикации: 21.03.2023

Аннотация. Проблема читаемости текста на естественном русском языке актуальна для людей с различными когнитивными нарушениями и для тех, кто слабо владеет языковыми знаниями, например, трудовых мигрантов и детей. Повышение доступности текстов (инструкций, указаний, рекомендаций) для указанных категорий граждан возможно путем использования автоматизированного алгоритма симплификации текста. В данном исследовании в качестве автоматизированного алгоритма симплификации используются глубокие нейронные архитектуры – трансформеры. В работе были применены следующие языковые модели: ruT5-base-absum, ruT5-base-paraphraser, ruT5_base_sum_gazeta, ruT5-base. Экспериментальные исследования проведены с использованием двух наборов данных – Института филологии и языковой коммуникации и из открытого репозитория Github. Для оценки моделей использован набор метрик: BLEU, индекс удобочитаемости Флеша, автоматический индекс удобочитаемости и разница длин предложений. С помощью тестового набора данных из перечисленных метрик извлекались статистические показатели, на основе которых сравнивались алгоритмы с различными параметрами обучения. Было проведено несколько экспериментов с указанными моделями, в которых использовались разные значения параметра скорости обучения для каждого набора данных, размеры батча, а также исключение из обучения дополнительного набора данных. Несмотря на различные показатели метрик при ручном сравнении выходы моделей слабо отличались друг от друга. Результаты экспериментальных исследований показали необходимость увеличения набора данных для обучения моделей, а также изменения параметров обучения моделей или использования других алгоритмов. Данное исследование является первым шагом к созданию системы поддержки принятия решений для автоматического упрощения текста и требует дальнейшего развития.
Ключевые слова: обработка естественного языка, симплификация текста, глубокое обучение, модель T5

Проблема преобразования сложных текстов в тексты на понятном языке имеет существенное социальное значение. С трудностями прочтения текстов, содержащих различные инструкции и правила, сталкиваются практически все. Проблема особенно актуальна для людей старшего возраста с первыми симптомами когнитивных нарушений, для мигрантов, знание русского языка которых не позволяет в достаточной степени понять текстовую информациею, людей, перенесших травмы головного мозга. Особенности восприятия всех этих групп людей не учитываются при написании текстов, ориентированных на широкий круг адресатов, например, в текстах социального взаимодействия (объявлений, информационных проспектов, размещаемых в поликлиниках, почтовых отделениях, филиалах пенсионного фонда, органах социальной защиты и т.д.).

Преобразование текста с целью удаления языковых конструкций, осложняющих его восприятие, называется симплификацией, или упрощением текста. Симплификация текста в автоматизированном режиме позволяет реализовать правила упрощения текстов с помощью инструментов обработки стандартного языка. С использованием механизма автоматизированной симплификации текста может быть обеспечен равный доступ к текстовой информации для различных категорий населения.

Постановка задачи и технологии

Упрощение текста направлено на преодоление трудностей его восприятия путем уменьшения лингвистической сложности без потери исходной информации и смысла. Уменьшение сложности текста и некоторые другие схожие задачи, такие как резюмирование и парафразирование, могут использоваться для сокращения времени и сил на чтение текстов, например, в Википедии на упрощенном английском языке – разделе Википедии для читателей, которым удобнее использовать упрощенный английский. Примеры фрагментов текста и их упрощенные варианты показаны в таблице 1. Упрощение выполнено экспертами *Института филологии и языковой коммуникации* (ИФиЯК) Сибирского федерального университета.

В показанных примерах для упрощения текста применены следующие приемы: удаление избыточных слов, преобразование структуры

Таблица 1

Примеры упрощения текста

Table 1

Text simplification examples

Исходный текст	Упрощенный текст
Аналогичные товары, а также серебро и ценные изделия из дерева, украшенные золотом и слоновой костью, египтяне получали из Вавилона.	Египтяне получали из Вавилона серебро, ценные изделия из дерева, украшенные золотом и слоновой костью.
Английский язык имеет многовековую историю становления, развития и территориального распространения, которая неразрывно связана с изменением языка, происходящим с течением времени, а также географическим и социальным разнообразием его употребления.	История английского языка связана с его изменением в течение времени, с географическим и социальным разнообразием.
На первом фестивале были показаны такие картины, как «Гедда» с Глендой Джексон, «Черная Луна» и «Призрак свободы». Третий фестиваль в 1978 году проводился под руководством Радживы Гупты.	Фильмы «Гедда», «Черная луна», «Призрак свободы» показывались на первом фестивале. Третий фестиваль проводил Раджив Гупта.
Парадоксально, но, обладая почти пятикратным перевесом и находясь в выгодной с тактической точки зрения обороне, китайцы проиграли битву.	Китайцы проиграли битву вопреки пятикратному перевесу и тактически выгодной обороне.
Рост сферы услуг по предоставлению помощи рекордному числу беженцев во время европейского миграционного кризиса способствовал снижению уровня безработицы внутри страны.	Помощь беженцам снизила уровень безработицы внутри страны.

предложений, изменение порядка слов. Использование этих приемов выполнено с учетом сохранения основного смысла предложения без потери существенных деталей.

Симплификация текста относится к области обработки естественного языка и актуальна при автоматизации задач, традиционно считающихся интеллектуальными: машинный перевод, морфологический анализ, распознавание текстов, преобразование звуковой информации в текст и др. [1].

Использование технологий глубокого обучения позволило добиться качественно лучших результатов работы алгоритма симплификации текстов. К одной из первых моделей относится DRESS [2], разработанная в 2017 году и имеющая в основе архитектуру рекуррентной сети LSTM, модель энкодер-декодер, механизм внимания и средства награды системного выхода. Такая модель при улучшении качества работы и увеличении независимости от языка текста также обладала значительными недостатками, обусловленными использованием рекуррентной сети. Рекуррентная модель требует учета состояний от текущего до всех предыдущих. Кроме того, в этой модели нет возможности учета единого контекста слов в текущей связке, поэтому в рекуррентных сетях для обработки текстов используют двойной контекст – прямой и обратный, другими словами, нейросеть видит либо все предыдущие слова в связке от текущего, либо следующие.

С 2018 года стали набирать популярность модели, основанные на архитектуре трансформера [3]. Использование таких моделей (рис. 1) для решения задач обработки естественного языка (машинный перевод, семантический анализ текстов и многие другие) показало качественно новый уровень результатов.

Трансформеры являются абсолютными лидерами в области обработки естественного языка, поскольку эффективно решают поставленные задачи, а благодаря доступности предобученных моделей автоматизируют все новые и новые задачи. Представленная модель содержит по два энкодера и декодера, но в общем случае их число может быть произвольным, однако должно быть одинаковым.

Наиболее популярной моделью при решении задачи обработки естественного языка является BERT [4]. Для решения задач, связанных с преобразованием одних текстов в другие, используют языковые модели T5 [5], которые также основаны на трансформере. Удачный опыт применения трансформеров для решения задач обработки текстов на русском языке показали модели T5 [6], RuGPT3 [7] и Mbart [8]. Модели трансформера успешно используются для задачи двойного машинного перевода: русский текст переводится на английский язык и обратно [9].

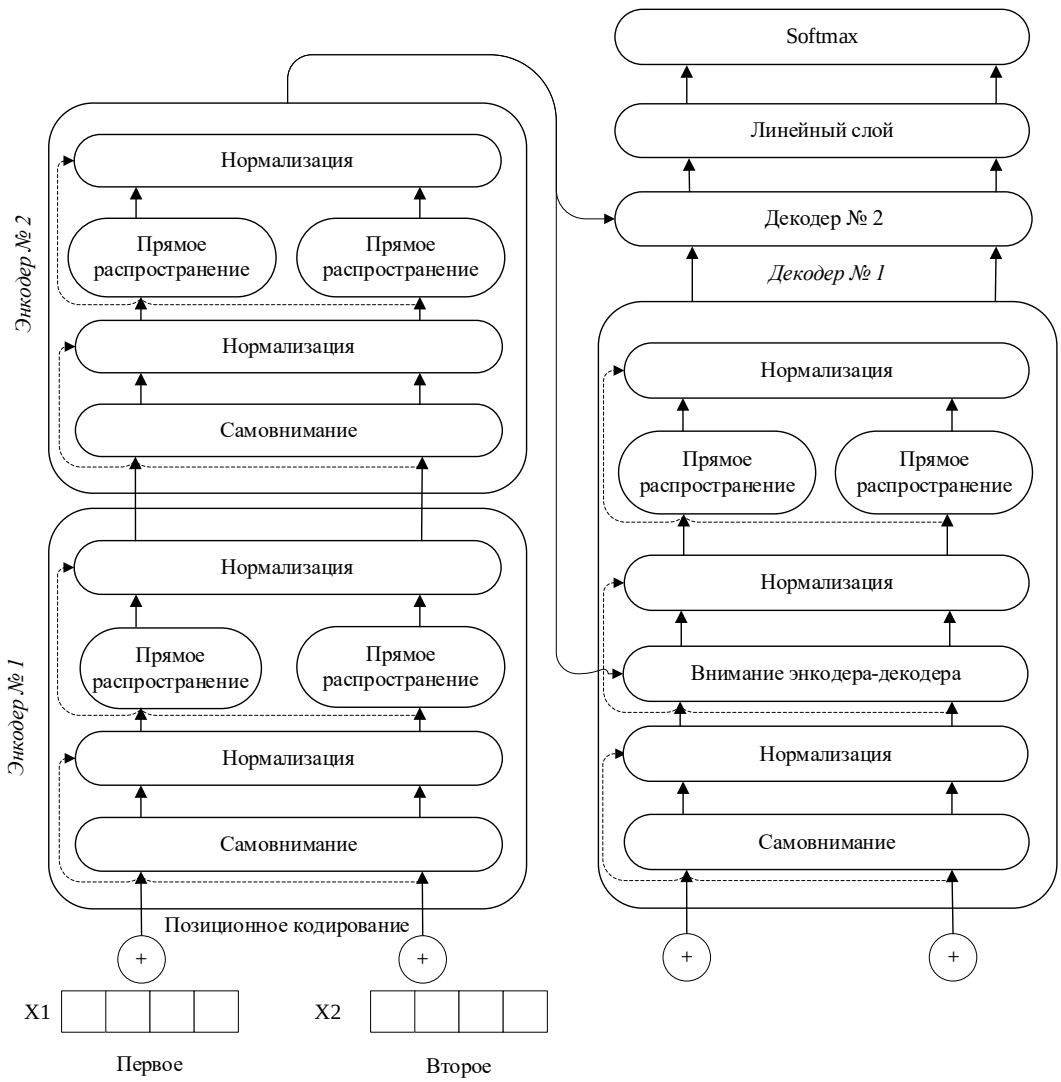

Рис. 1. Архитектура языковой модели на базе трансформера

Fig. 1. Architecture of the transformer language model

Перечисленные модели имеют значительные отличия. Так, в модели BERT используются только энкодеры, можно работать с текстом в прямом и обратном направлениях, с методом обучения и некоторыми другими небольшими деталями. Модели GPT [10], наоборот, используют стек декодеров, а также другой тип самовнимания. В модели T5 входящая и выходящая информация является строками, ее архитектура отличается изменением слоя нормализации, в котором выполнено удаление смещения слоя, а также размещения слоя за пределами остаточного пути. Кроме того, в T5 применяется другая схема кодирования позиции токенов.

Эти изменения позволили использовать перенос обучения, что обусловило повышение эффективности работы и уменьшение затрат на обучение.

Применение модели T5 для симплификации текста

Перед решением задачи обработки текста необходимо представить его в удобном для алгоритмов виде. Для более ранних алгоритмов предобработки текста использовалось множество шагов, таких как сегментация текста на предложения, перевод его в нижний регистр, удаление или замена знаков пунктуации и/или чисел, перевод слов в начальную форму, токенизация, удаление специальных стоп-слов. Для предобученных моделей этот этап может отсутствовать полностью, так как современные

языковые модели используют всю информацию в тексте. Однако в некоторых случаях он остается: например, если набор данных собран автоматически из отзывов на товары, можно добавить фильтры удаления смайликов, эмодзи, ссылок, оскорбительных слов и выражений. В данной работе предобработка текста заключается в приведении его в одну кодировку.

Применение моделей трансформера для обработки текста выполняется в несколько этапов: кодирование текста в машиночитаемый вид, прохождение слоев нейронной сети и перевод обратно в человекочитаемый вид.

Этап 1. Текст конвертируется в числовой вид, токенизируется, при этом токенами могут являться как отдельные символы, так и части слова и даже целые слова. Для токенизации наиболее часто применяется метод BPE, использующий части слов и сжатие данных. Он гарантирует, что наиболее распространенные слова представлены в словарном запасе как один токен, в то время как редкие слова разбиваются на два или более токенов. Кроме самих слов, кодируются и позиции токенов в предложении. Преобразованный таким образом текст в машинном представлении поступает в модель для дальнейшей обработки.

Этап 2. Токенизированный текст последовательно проходит через стек энкодеров. Энкодер получает на вход список векторов (тензоров), обрабатывает его, передавая векторы в слой самовнимания, затем после нормализации передает данные в нейронную сеть с прямой связью, а далее отправляет их к следующему энкодеру. Нормализация состоит в масштабировании входных значений тензора таким образом, чтобы среднее значение было нулем, а дисперсия равнялась единице.

Этап 3. После прохождения стека энкодеров текст передается в стек декодеров, устройство которых близко к энкодеру (рис. 1). После слоя самовнимания и перед подачей данных в сети с прямым распространением используется дополнительный механизм внимания энкодера-декодера. Данный слой имеет отличия от механизма самовнимания, так как может фокусироваться только на предыдущих позициях в выходном предложении (это выполняется с помощью маскировки всех позиций после текущей). Кроме того, декодер создает матрицу запроса из слоя, который находится ниже, и берет матрицы ключей (Keys) и значений (Values) из выхода стека энкодеров. Таким образом, выходные наборы векторов K и V верхнего энкодера используются всеми декодерами в механизме внимания энкодера-декодера.

Этап 4. Происходит детокенизация, или расшифровка векторных значений обратно в слова. Стек декодеров возвращает вектор чисел, который проходит через полносвязный линейный слой. Выходом линейного слоя являются так называемые логит-векторы – матрицы, имеющие размерность, равную числу слов в словаре, и количество, равное выходным токенам. Логит-векторы представляют собой распределенную вероятность выходного слова, и, чтобы вычислить из векторов самое вероятное слово, необходимо применить функцию активации softmax. После функции активации векторы заменяются конкретным словом, а матрицы преобразуются в список выходных слов.

Таким образом, обработка текста на основе модели трансформера заключается в выполнении следующих шагов: токенизация исходного текста, прохождение текста в машинном виде последовательно через стек энкодеров и декодеров и преобразование его обратно в текст (рис. 2).

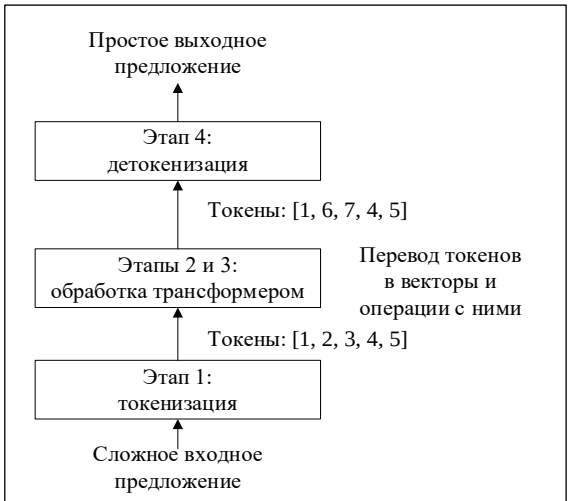

Рис. 2. Алгоритм симплификации текста

Fig. 2. A text simplification algorithm

Механизм внимания можно описать как алгоритм, помогающий фокусироваться на некоторых (предположительно, более важных) областях данных. Он используется для исключения потери смысла предложений при решении задачи симплификации текста. Существует несколько видов механизмов внимания. Первоначально для обработки языка использовался механизм внимания, напрямую сравнивающий два разных предложения. Однако после появ-

ления трансформеров используется модификация этого механизма, названная механизмом самовнимания, который строит матрицу отношений между токенами для предложений. Сравнивая токены в одном предложении между собой, этот механизм создает их контекст.

Механизм самовнимания работает следующим образом: выполняется сопоставление запроса и набора пар ключ–значение с выходными данными, при этом запрос, ключи, значения и выходные данные являются векторами. Результат работы вычисляется как взвешенная сумма значений, вес, присвоенный каждому значению, рассчитывается как функция совместимости запроса с соответствующим ключом: $Attention(Q, K, V) = \text{softmax}(QK^T/DimK)V$, где Q – вектор запроса; K – вектор ключей; V – вектор значений; K^T – транспонированный вектор ключей; $DimK$ – квадратный корень размерности вектора.

Такой механизм самовнимания называется механизмом одноголового внимания и является неэффективным во многих отношениях по сравнению с многоголовым вниманием. Для увеличения скорости работы за счет распараллеливания вычислений, увеличения точности и большей устойчивости к ошибкам используют несколько голов внимания. При этом пространство представления токенов разбивается на проекции или подпространства и каждой голове присваивается свое подпространство. Результатом вычисления мультиголового внимания является последовательная конкатенация выходных матриц каждой головы, что можно выразить следующей формулой: $MultiHead(Q, K, V) = Concat(head_1, \ldots, head_N)W^o$, где Q – вектор запроса; K – вектор ключей; V – вектор значений; $head_i$ – выход i-й головы; N – количество голов; W^o – весовая матрица отображения из пространства запроса и пространства ключей в пространство значений.

Разбив пространство токенов на подпространства (проекции), можно добиться улучшения качества работы модели за счет учета деталей контекста, а также увеличить скорость обработки данных за счет распараллеливания вычислений, которые иначе применялись бы к векторам и матрицам большей размерности.

Процесс обработки текста при реализации алгоритма симплификации показан на рисунке 3.

Из тестового набора данных извлекаются предложения, на каждое из которых модель генерирует ответ (предположительно, более простой). Далее из начальных предложений и от

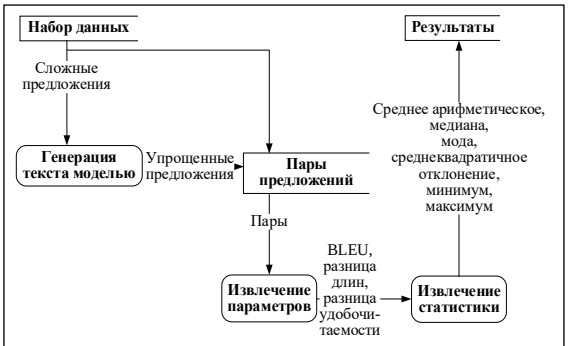

Рис. 3. Алгоритм оценки работы языковых моделей

Fig. 3. An algorithm for evaluating work of language models

ветов составляются пары предложений. Каждая тысяча пар предложений формируется в таблицу, из которой извлекаются параметры для каждой пары. Далее на основе извлеченных параметров формируется статистика. Результатом анализа является таблица статистических данных по параметрам оценки модели.

Экспериментальные исследования

Для обучения моделей, решающих задачи симплификации текста, использованы набор данных ИФиЯК и данные из открытого репозитория Github. Набор данных ИФиЯК (https://drive.google.com/file/d/1AxJt-Z_0utMqCLzQZ45UIClPBjqj0Mud/view?usp=sharing) составлен из текстов информационных объявлений и состоит из пар исходных и упрощенных экспертами-лингвистами предложений. Набор разбит на две части, 195 и 5 пар предложений. Первая часть использовалась для обучения, вторая – для проверки экспертом. Дополнительный набор данных получен из репозитория GitHub (https://github.com/dialogue-evaluation/RuSimpleSentEval). Он собран организаторами Международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 21» (https://www.dialog-21.ru/) и представляет собой CSV-файл, содержащий 3 406 пар «сложное–простое предложение». Особенностью этого набора данных является наличие нескольких различных вариантов простых предложений на каждое сложное.

Тестирование моделей выполнено с помощью 1 000 сложных предложений из набора данных репозитория (указан в репозитории как файл public test). Сложные предложения поступали на вход обученной модели, и к ним гене-

рировался ответ. На основе сравнения исходного и полученного предложений делался вывод о качестве работы модели. Сравнение качества работы модели выполнено с использованием разницы длины текстов и метрик FRE, ARI, BLEU.

Разница длины генерируемого и оригинального текстов использована для оценки потери информации. Отрицательные числа означают сокращение количества слов и косвенно потерю информации. Следует отметить, что сокращение на 1–3 слова в среднем не мешает восприятию смысла текста.

Метрика FRE – это индекс удобочитаемости текста Флеша, рассчитываемый по формуле

$$FRE = 206{,}835 - 1{,}52ASL - 65{,}14ASW, \quad (1)$$

где *ASL* – средняя длина предложения в словах; *ASW* – средняя длина слова в слогах.

Под удобочитаемостью будем понимать вариант английского термина readability [11] – сложную взаимосвязь между текстом и усилием среднестатистического читателя, затраченным на его прочтение. Для расчета удобочитаемости методов FRE и ARI использован фреймворк ruTS (https://github.com/SergeyShk/ruTS). Значения FRE представляют собой вещественную оценку в диапазоне от 0 до 100, где 100 – самый высокий показатель удобочитаемости, увеличение индекса означает улучшение удобочитаемости по сравнению с оригиналом.

Автоматический индекс удобочитаемости ARI вычисляется согласно выражению

$$ARI = 4{,}71(C/W) + 0{,}5(W/S) - 21{,}43, \quad (2)$$

где *C* – количество букв и цифр в тексте; *W* – количество слов в тексте; *S* – количество предложений в тексте. Метрика удобочитаемости ARI имеет значения в диапазоне от 1 до 14, где 1 означает самый высокий показатель удобочитаемости.

Метрики FRE и ARI не учитывают правила построения предложений, слов, грамматику и сложные зависимости, соответственно, не могут целиком заменить оценку человеком. В силу этого для оценки качества моделей требуются дополнительные специальные метрики. Одной из популярных метрик, имеющих высокую корреляцию с оценкой человека, является метрика SARI [12], однако она требует использования кортежа из трех предложений, а в наборах данных предложения расположены парами. В связи с этим в настоящей работе использована метрика BLEU [13], которая разрабатывалась для оценки качества работы систем машинного перевода. Значение метрики лежит в интервале от 0 до 1, где 0 означает полное различие текстов, а 1 – идентичность, то есть с помощью этой метрики оценивается корреляция исходного текста и выхода модели. Диапазон значений BLEU от 0,8 до 1 означает совсем незначительные изменения исходного текста, от 0,5 до 0,8 – приемлемые, ниже – слишком значительные.

Таким образом, для оценки моделей использованы набор метрик BLEU, разница FRE (1), разница ARI (2) и разница длины предложений. Под разницей подразумевается параметр, рассчитанный для выходного текста модели, минус параметр, рассчитанный для исходного текста. По выбранным метрикам рассчитаны минимальное, максимальное, медианное и среднее арифметические значения, мода и среднеквадратичное отклонение.

Для эксперимента были использованы следующие модели на базе архитектуры T5.

- Absum (https://huggingface.co/cointegrated/rut5-base-absum) – модель для задачи абстрактного резюмирования текста.
- Paraphraser (https://huggingface.co/cointegrated/rut5-base-paraphraser) – модель, обученная на задаче перефразирования фрагментов текста.
- Reference – модель, идентичная Paraphraser (эта модель ни в одном эксперименте не была подвергнута дообучению).
- Sum Gazeta (https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta) – модель, обученная на газетных новостях и их резюме.
- Base (https://huggingface.co/sberbank-ai/ruT5-base) – модель, основанная на оригинальном коде от Google, обученная на 300 Гб текстовых данных.

В таблице 2 показаны результаты работы моделей T5 с коэффициентом скорости обучения 1E–4 и размером батча 5. Обучение происходило сначала на дополнительном наборе, а затем на целевом. Именно эти характеристики оказались лучшими при решении задачи симплификации текста. Символом $\bar{X}$ обозначено среднее значение метрики, СКО – среднеквадратичное отклонение.

Полужирным текстом выделены минимальные значения. Для сравнения эффективности параметры оценки были извлечены из основного набора данных (набор пар: исходное сложное предложение, упрощенное лингвистами предложение). В таблице 2 указывается абсолютная разница между целевым значением параметра (извлеченным из целевого набора) и вычисленным в ходе эксперимента, выраженная в процентах.

Примеры пар предложений для модели Paraphraser с параметрами, указанными в таблице 2, отражены в таблице 3.

Модель изменила входные предложения только в двух из пяти случаев, при этом удалила часть текста в качестве изменения. Похожая картина наблюдается и в других моделях. Ни одна модель полностью не справилась с поставленной задачей, что говорит о необходимости доработки и увеличении собственного набора данных с целью повышения эффективности работы алгоритма автоматизированной симплификации текста. Лучшее качество работы алгоритма упрощения текста показала модель Reference, которая не была дополнительно обучена.

Заключение

Таким образом, симплификация текста является сложной задачей, имеющей высокую социальную значимость. Для выполнения симплификации в автоматизированном режиме в работе использованы трансформеры типа T5. Качество работы модели оценивалось с использованием интегральной характеристики с учетом специфических метрик оценки текстов, а также экспертами-лингвистами. В результате исследования был сделан вывод о недостаточном качестве работы алгоритма автоматической симплификации. Самой вероятной причиной этого является недостаточное количество обучающих данных. Авторы планируют даль-

Таблица 2

Результаты эксперимента (в процентах)

Table 2

Experimental results (in percentage terms)

Параметр	Reference		Absum		Paraphraser		Sum Gazeta		Base	
	$\bar{X}$	СКО	$\bar{X}$	СКО	$\bar{X}$	СКО	$\bar{X}$	СКО	$\bar{X}$	СКО
BLEU	**7,0**	**0,3**	23,5	3,9	21,6	5,3	21,7	1,9	20,4	3,9
Разница длины	26,2	60,4	42,9	117,6	**0,1**	**19,2**	93,5	152,9	41,4	40,2
Разница FRE	571,8	**655,3**	511,0	888,7	519,4	934,0	**455,7**	776,6	481,7	882,8
Разница ARI	100,4	**117,5**	99,6	161,4	98,7	172,7	88,3	140,1	**83,6**	168,0

Таблица 3

Пары предложений модели Paraphraser

Table 3

Paraphraser model pairs of sentences

Исходный текст	Выход модели Paraphraser
Врач соберет необходимые сведения о болезни и факторах, которые могут служить причиной появления аллергической реакции, назначит необходимые исследования.	Врач соберет необходимые сведения о болезни и факторах, которые могут служить причиной появления аллергической реакции.
Категорически запрещается подвергать карту механическим воздействиям, которые могут даже незначительно повредить ее.	Категорически запрещается подвергать карту механическим воздействиям, которые могут даже незначительно повредить ее.
Направление на данное исследование выдает только врач-аллерголог (иммунолог), который обязан проинформировать вас о том, в каких именно медицинских организациях вы можете провести это исследование бесплатно.	Направление на данное исследование выдает только врач-аллерголог (иммунолог), который обязан проинформировать вас о том, в каких именно медицинских организациях вы можете провести это исследование бесплатно.
Размер страховой пенсии по случаю потери кормильца подлежит пересчету с 1 августа года, следующего за годом, в котором была назначена эта пенсия.	Размер страховой пенсии по случаю потери кормильца подлежит пересчету с 1 августа.
Проверить, какие льготы установлены в конкретном регионе или муниципальном образовании, можно на сайте ФНС.	Проверить, какие льготы установлены в конкретном регионе или муниципальном образовании, можно на сайте ФНС.

нейшую доработку языковых моделей для решения задачи симплификации текста, увеличение набора данных путем самостоятельного сбора текстовой информации в учреждениях социальной направленности в городе Красноярске и с сайтов таких учреждений по всей России. Далее в сотрудничестве с экспертами запланирована обработка этих текстов с преобразованием на простой язык и формированием пар предложений. Полученный набор данных будет использован для обучения и тестирования модели. Разработанный метод станет основным для программной разработки с веб-интерфейсом, которая позволит преобразовывать сложный текст с камеры мобильного телефона в режиме реального времени, что будет способствовать повышению доступности текстовых материалов для различных категорий граждан.

Список литературы

1. Васильев Д.Д., Пятаева А.В. Модель представления языка Bert // РИИ: матер. XIV науч.-технич. конф. 2022. С. 94–97.
2. Zhang X., Lapata M. Sentence simplification with deep reinforcement learning. Proc. Conf. on Empirical Methods in Natural Language Processing, 2017, pp. 584–594. doi: 10.18653/v1/D17-1062.
3. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L. et al. Attention is all you need. NIPS, 2017, pp. 5998–6008.
4. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. NAACL-HLT, 2019, pp. 4171–4186.
5. José M., Micaela A., Sílvia A. Using a pre-trained simpleT5 model for text simplification in a limited corpus. CEUR-WS Proc., 2022, pp. 1–6.
6. Raffel C., Shazeer N., Roberts A., Lee K., Narang S. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J. of Machine Learning Research, 2019, vol. 21, pp. 5485–5551.
7. Fenogenova A. Text simplification with autoregressive models. Proc. Computational Linguistics and Intellectual Tech., 2021, pp. 1–8. doi: 10.28995/2075-7182-2021-20-227-234.
8. Liu Y., Gu J., Goyal N., Li X., Edunov S. et al. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 2020, vol. 8, pp. 726–742. doi: 10.1162/tacl_a_00343.
9. Galeev F., Leushina M., Ivanov V. ruBTS: Russian sentence simplification using back-translation. Proc. Computational Linguistics and Intellectual Tech., 2021, pp. 1–8. doi: 10.28995/2075-7182-2021-20-259-267.
10. Brown T.B., Mann B., Ryder N. et al. Language models are few-shot learners. Proc. NeurIPS, 2020, pp. 1877–1901.
11. Солнышкина С.И., Кисельников А.С. Сложность текста: этапы изучения в отечественном прикладном языкознании // Вестн. ТГУ. Филология. 2015. № 6. С. 86–99.
12. Xu W., Napoles C., Pavlick E., Chen Q., Callison-Burch. C. Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics, 2016, vol. 4, pp. 401–415. doi: 10.1162/tacl_a_00107.
13. Wołk K., Koržinek D. Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking. Comput. Sci., 2015, vol. 18, no. 2, pp. 129–144. doi: 10.7494/csci.2017.18.2.129.

Software & Systems doi: 10.15827/0236-235X.142.228-236 2023, vol. 36, no. 2, pp. 228–236

T5 language models for text simplification

D.D. Vasiliev
A.V. Pyataeva

For citation
Vasiliev, D.D., Pyataeva, A.V. (2023) 'T5 language models for text simplification', *Software & Systems*, 36(2), pp. 228–236 (in Russ.). doi: 10.15827/0236-235X.142.228-236
Article info
Received: 13.03.2023 After revision: 14.03.2023 Accepted: 21.03.2023

Abstract. The problem of text readability in natural Russian is relevant for people with various cognitive impairments and for people with poor language skills, such as labor migrants or children. Texts constantly surround us in real life, such as

various instructions, directions, and recommendations. Increasing the availability of these texts for these categories of citizens is possible by using an automated text simplification algorithm. This article used deep neural architecture transformers as an automated simplification algorithm. The following language models were applied: ruT5-base-absum, ruT5-base-paraphraser, ruT5_base_sum_gazeta, ruT5-base. Experimental studies used two data sets: a data set from the Institute of Philology and Language Communication and data from the open Github repository. The following set of metrics was used to evaluate the models: BLEU, Flesh Readability Index, Automatic Readability Index, and Sentence Length Difference. Further, using a test data set, statistical indicators were extracted from the listed metrics, which became the basis for comparing algorithms with different training parameters. The authors carried out several experiments with these models that used different values of the learning rate parameter for each dataset, batch sizes, and the exclusion of an additional dataset from training. Despite the different metrics, the models outputs did not differ much from each other during manual comparison. The results of experimental studies show the need to increase the data set for model training, as well as the change in the parameters of model training, or the use other algorithms. This study is the first step towards creating a decision support system for automatic text simplification and requires further development.
Keywords: natural language processing, text simplification, deep learning, T5 model

Reference List

1. Vasiliev, D.D., Pyataeva, A.V. (2022) 'The Bert language representation model', *Proc. Conf. Robotics and Artificial Intelligence*, pp. 94–97 (in Russ.).
2. Zhang, X., Lapata, M. (2017) 'Sentence simplification with deep reinforcement learning', *Proc. Conf. on Empirical Methods in Natural Language Processing*, pp. 584–594. doi: 10.18653/v1/D17-1062.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. et al. (2017) 'Attention is all you need', *NIPS*, pp. 5998–6008.
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *Proc. NAACL-HLT*, pp. 4171–4186.
5. José, M., Micaela, A., Sílvia, A. (2022) 'Using a pre-trained simpleT5 model for text simplification in a limited corpus', *CEUR-WS Proc.*, pp. 1–6.
6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S. et al. (2019) 'Exploring the limits of transfer learning with a unified text-to-text transformer', *J. of Machine Learning Research*, 21, pp. 5485–5551.
7. Fenogenova, A. (2021) 'Text simplification with autoregressive models', *Proc. Computational Linguistics and Intellectual Tech.*, pp. 1–8. doi: 10.28995/2075-7182-2021-20-227-234.
8. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S. et al. (2020) 'Multilingual denoising pre-training for neural machine translation', *Transactions of the Association for Computational Linguistics*, 8, pp. 726–742. doi: 10.1162/tacl_a_00343.
9. Galeev, F., Leushina, M., Ivanov, V. (2021) 'uBTS: Russian sentence simplification using back-translation', *Proc. Computational Linguistics and Intellectual Tech.*, pp. 1–8. doi: 10.28995/2075-7182-2021-20-259-267.
10. Brown, T.B., Mann, B., Ryder, N. et al. (2020) 'Language models are few-shot learners', *Proc. NeurIPS*, pp. 1877–1901.
11. Solnyshkina, M.I., Kiselnikov, A.S. (2015) 'Text complexity: Study phases in Russian linguistics', *Tomsk State University J. of Philology*, (6), pp. 86–99 (in Russ.).
12. Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C. (2016) 'Optimizing statistical machine translation for text simplification', *Transactions of the Association for Computational Linguistics*, 4, pp. 401–415. doi: 10.1162/tacl_a_00107.
13. Wołk, K., Koržinek, D. (2015) 'Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking', *Comput. Sci*., 18(2), pp. 129–144. doi: 10.7494/csci.2017.18.2.129.

Авторы
Васильев Дмитрий Дмитриевич [1], магистрант, dmitriy.vasiliev.0303@gmail.com
Пятаева Анна Владимировна [1], к.т.н., доцент, anna4u@list.ru

[1] Сибирский федеральный университет, г. Красноярск, 660074, Россия

Authors
Dmitriy N. Vasiliev [1], Graduate Student, dmitriy.vasiliev.0303@gmail.com
Anna V. Pyataeva [1], Ph.D. (Engineering), Associate Professor, anna4u@list.ru

[1] Siberian Federal University, Krasnoyarsk, 660074, Russian Federation