MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 31, NO. 2, 2024
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Keywords, Morpheme Parsing and Syntactic Trees: Features for Text Complexity Assessment

D. A. Morozov[1], I. A. Smal[1], T. A. Garipov[1], A. V. Glazkova[2]

DOI: 10.18255/1818-1015-2024-2-206-220

[1]Novosibirsk National Research State University, Novosibirsk, Russia
[2]University of Tyumen, Tyumen, Russia

MSC2020: 68T50
Research article
Full text in Russian

Received February 27, 2024
Revised March 29, 2024
Accepted May 8, 2024

The text complexity assessment is an applied problem of current interest with potential application in the drafting of legal documents, editing textbooks, and selecting books for extracurricular reading. The methods for generating a feature vector when automatically assessing the text complexity are quite diverse. Early approaches relied on easily calculable quantities, such as the average length of a sentence or the average number of syllables per word. With the development of natural language processing algorithms, the space of used features is expanding. In this work, we examined three groups of features: 1) automatically generated keywords, 2) information about the features of morphemic word parsing, and 3) information about the diversity, branching, and depth of syntactic trees. The RuTermExtract algorithm was utilized to generate keywords, a convolutional neural network model was used to generate morphemic parses, and the Stanza model, trained on the SynTagRus corpus, was used to generate syntax trees. We conducted a comparison using four different machine learning algorithms and four annotated Russian-language text corpora. The corpora used differ both in the domain and markup paradigm, due to which the results obtained more objectively reflect the real relationship between the characteristics and the text complexity. The use of keywords performed worse on average than the use of topic markers obtained using latent Dirichlet allocation. In most situations, morphemic characteristics turned out to be more effective than previously described methods for assessing the lexical complexity of a text: the frequency of words and the occurrence of word-formation patterns. The use of an extensive set of syntactic features allowed, in most cases, to improve the quality of work of neural network models in comparison with the previously described set.

Keywords: text complexity; keyword generation; morpheme parsing generation; syntax trees

INFORMATION ABOUT THE AUTHORS

Morozov, Dmitry A. (corresponding author)	ORCID iD: 0000-0003-4464-1355. E-mail: morozowdm@gmail.com Junior Researcher
Smal, Ivan A.	ORCID iD: 0009-0005-1082-0584. E-mail: vanasmal@mail.ru Graduate Student
Garipov, Timur A.	ORCID iD: 0009-0008-4527-2268. E-mail: garipov154@yandex.ru Master's Student
Glazkova, Anna V.	ORCID iD: 0000-0001-8409-6457. E-mail: a.v.glazkova@utmn.ru Associate Professor, PhD

For citation: D. A. Morozov, I. A. Smal, T. A. Garipov, and A. V. Glazkova, "Keywords, morpheme parsing and syntactic trees: features for text complexity assessment", *Modeling and Analysis of Information Systems*, vol. 31, no. 2, pp. 206–220, 2024. DOI: 10.18255/1818-1015-2024-2-206-220.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 31, № 2, 2024
сайт журнала: www.mais-journal.ru

Ключевые слова, морфемные разборы и синтаксические деревья в задаче оценки сложности текста

Д. А. Морозов[1], И. А. Смаль[1], Т. А. Гарипов[1], А. В. Глазкова[2] DOI: 10.18255/1818-1015-2024-2-206-220

[1]Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2]Тюменский государственный университет, Тюмень, Россия

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 27 февраля 2024 г.
После доработки 29 марта 2024 г.
Принята к публикации 8 мая 2024 г.

Задача оценки сложности текста является актуальной прикладной задачей с потенциальным применением при составлении юридических документов, редактуре учебников и подборе книг для внеклассного чтения. Способы формирования признакового описания при автоматической оценке сложности текста достаточно разнообразны. Ранние подходы опирались на легко вычислимые величины, такие как средняя длина предложения или среднее число слогов в слове. С развитием алгоритмов обработки естественного языка расширяется и пространство используемых признаков. В рамках настоящей работы мы исследовали три группы признаков: 1) автоматически генерируемые ключевые слова, 2) сведения об особенностях морфемного разбора слов и 3) информацию о разнообразии, разветвлённости и глубине синтаксических деревьев. Для генерации ключевых слов использован алгоритм RuTermExtract, для генерации морфемных разборов — свёрточная нейросетевая модель, для генерации синтаксических деревьев — модель Stanza, обученная на корпусе SynTagRus. Мы провели сравнение на материале четырёх различных моделей машинного обучения и четырёх аннотированных русскоязычных корпусов текстов. Использованные корпусы различаются как по домену, так и по парадигме разметки, благодаря чему полученные результаты объективнее отражают реальную связь характеристик и сложности текста. Использование ключевые слова показало в среднем результат хуже, чем использование тематических маркеров, получаемых при помощи латентного размещения Дирихле. Морфемные характеристики оказались в большинстве ситуаций эффективнее ранее описанных способов оценки лексической сложности текста: учёта частотности слов и встречаемости словообразовательных паттернов. Использование обширного набора синтаксических признаков позволило в большинстве случаев улучшить качество работы нейросетевых моделей в сравнении с ранее описанным набором.

Ключевые слова: сложность текста; генерация ключевых слов; генерация морфемных разборов; синтаксические деревья

ИНФОРМАЦИЯ ОБ АВТОРАХ

Морозов, Дмитрий Алексеевич (автор для корреспонденции)	ORCID iD: 0000-0003-4464-1355. E-mail: morozowdm@gmail.com Младший научный сотрудник
Смаль, Иван Андреевич	ORCID iD: 0009-0005-1082-0584. E-mail: vanasmal@mail.ru Аспирант
Гарипов, Тимур Александрович	ORCID iD: 0009-0008-4527-2268. E-mail: garipov154@yandex.ru Магистрант
Глазкова, Анна Валерьевна	ORCID iD: 0000-0001-8409-6457. E-mail: a.v.glazkova@utmn.ru Доцент, канд. тех. наук

Для цитирования: D. A. Morozov, I. A. Smal, T. A. Garipov, and A. V. Glazkova, “Keywords, morpheme parsing and syntactic trees: features for text complexity assessment”, *Modeling and Analysis of Information Systems*, vol. 31, no. 2, pp. 206–220, 2024. DOI: 10.18255/1818-1015-2024-2-206-220.

Введение

Сложность текста — величина, оценивать которую требуется при подготовке различных учебных пособий, инструкций, типовых договоров. При этом как экспертная, так и эмпирическая оценка этой величины требует значительных затрат времени и, зачастую, финансов. В связи с этим востребованными являются алгоритмы, позволяющие оценить сложность текста, исходя из вычисляемых (желательно автоматически) характеристик. Первые подобные алгоритмы были предложены в середине двадцатого века. Обычно они являлись линейными регрессиями, опиравшимися на простые статистические признаки, такие как средняя длина слов и предложений. К таким алгоритмам можно отнести индексы удобочитаемости, в частности, индекс Флеша [1], индекс Дейла — Чалла [2], Automated Readability Index [3]. При разработке и использовании алгоритмов оценки сложность текста обычно ассоциируют с возрастом предполагаемого читателя, а оценка проводится в терминах классов школы. При таком подходе читательский опыт носителей языка одного возраста считается слабо различающимся, а шкала ограничивается сверху старшими классами школы или младшими курсами университетов. Это позволяет использовать в качестве источника текстов и разметки различные учебные материалы и списки литературы.

Благодаря появлению всё более производительных компьютеров и развитию методов обработки естественного языка стало возможным улучшение качества оценки с использованием более продвинутых моделей и более сложных, высокоабстрактных и лингвистически мотивированных характеристик текста. Упоминаемые в различных работах признаки чрезвычайно разнообразны. В исследованиях, посвящённых оценке сложности русскоязычных текстов, помимо традиционно используемых признаков, таких как средние длины слов и предложений, встречается использование оценок лексического разнообразия [4—6], морфологических [5—9], синтаксических [6, 8—10], лексических [6—9, 11, 12], семантических [9], тематических [6] признаков.

Для оценки лексической сложности текста обычно используют характеристики, связанные с частотностью слов [6, 9, 12, 13]. При этом ряд исследований показал, что частотность, вычисленная на основании какого-либо корпуса, достаточно часто отличается от реальной распространённости и знакомости слов [11, 13]. В то же время, алгоритмы оценки именно сложности слов отсутствуют, а само понятие семантической сложности слова недостаточно формализовано. В этой ситуации способом оценить лексическую сложность текста целиком может стать вычисление различных особенностей морфемного разбора слов[1]. Так, в работе [9] используется информация о встречаемости различных ассоциированных со сложностью морфемных конструкций, а именно доли лемм вида **ция*, **ние*, **вие*, **тие*, **ист*, **изм*, **ура*, **ище*, **ство*, **ость*, **овка*, **атор*, **итор*, **тель*, **льный*, **овать*. Такой подход вычислительно проще полноценного анализа морфемных разборов, однако остаётся непонятным, насколько вычисленные таким образом признаки значимы.

В исследованиях, затрагивающих использование синтаксических признаков в рамках оценки сложности текста на русском языке [6, 8—10], набор задействованных признаков достаточно узок. Чаще всего авторы используют глубину синтаксического дерева, долю рёбер того или иного типа и расстояние между вершиной и её непосредственным предком, выбор используемых признаков при этом как правило интуитивен. При этом в работе [10] синтаксические признаки оказались среди наиболее значимых как в отдельных корпусах, так и в целом. В работе [6] использование синтаксических признаков позволило улучшить качество модели для большинства корпусов и алгоритмов машинного обучения. Таким образом, представляется интересным изучение более обширных наборов синтаксических признаков, а также поиск наиболее значимых из них.

[1]Морфемным разбором слова называют разбиение слова на несколько непересекающихся подстрок-морфем (возможно, с добавлением дополнительных символов), которые делятся на различные классы согласно своей природе: корни и разнообразные аффиксы (приставки, суффиксы, окончания и т. д.).

Информация о тематике текста относительно редко применяется при оценке сложности текста. В то же время, в работе [6] использование латентного размещения Дирихле (LDA) [14] для формирования тематических кластеров позволило значительно (на 3–10 %) повысить качество работы модели. Это позволяет предположить, что другие механизмы извлечения терминов, описывающих тематику текста, могли бы быть использованы в качестве признаков. Одним из способов кратко описать тематику текста является использование ключевых слов. Ключевые слова представляют собой набор слов и словосочетаний, в совокупности соответствующих высокоуровневому описанию содержания текста. Они широко применяются в научной среде и СМИ, но в общем случае, особенно для длинных художественных текстов, этот инструмент, по-видимому, либо мало применим, либо неприменим вовсе из-за разнообразия обсуждаемых в тексте тем. Однако в случае с оценкой сложности текста многие из исследований проводятся на корпусах, состоящих из небольших отрывков текста. Так, в работе [10] в том числе используются фрагменты по 200 предложений, в работе [6] — по 70 предложений, в работе [9] — 11 ± 7 предложений. При такой длине контекста, сравнимой с длиной аннотаций к научным статьям или газетных текстов[2], генерируемые ключевые слова, вероятно, могут быть использованы в качестве признаков тематического моделирования при оценке сложности текста и сравнены с ранее исследованным LDA.

Итак, целью настоящей работы является анализ возможных расширений признакового описания в задаче автоматической оценки сложности текста в терминах возраста его предполагаемого читателя. Мы рассмотрели три направления таких расширений: добавление признаков, связанных с лексической сложностью, с синтаксической сложностью и с тематикой текста. Для каждого из направлений мы сравнили влияние на качество оценки ранее упоминавшихся признаков с вновь предлагаемыми.

Статья организована следующим образом. В разделе 1 описаны использованные при сравнении значимости корпусы текстов и модели машинного обучения. Раздел 2 состоит из трёх подразделов, в каждом из которых описывается методика извлечения отдельного типа признаков. Подраздел 2.1 содержит описание вычисления ключевых слов и построения на их базе признаков. В подразделе 2.2 описано построение модели морфемных разборов и её применение для извлечения характеристик текста. Наконец, подраздел 2.3 содержит описание экспериментов, направленных на поиск наиболее значимых особенностей синтаксических деревьев. В разделе 3 приводятся и обсуждаются полученные результаты. Выводы к работе представлены в заключении.

1. Модели и данные

Для того, чтобы иметь возможность более прозрачно сравнить вновь полученные результаты с имевшимися ранее, для проверки признаков мы выбрали наборы данных из числа ранее упоминавшихся в контексте оценки сложности текста. Мы использовали четыре корпуса русскоязычных текстов: корпус ознакомительных отрывков (Fic), корпус учебных текстов (TB), корпус рекомендованной литературы (RL) и корпус читаемых книг (BR). Во всех этих корпусах в качестве разметки сложности выступает разметка возраста читателя. Корпус Fic был представлен в работе [15] и состоит из начальных фрагментов различных произведений, открыто доступных для ознакомления в онлайн-библиотеках. Для этого корпуса использовалось два вида имеющейся в нём разметки — а) двухклассовая разметка на тексты для детей и для взрослых (FicChAd) и б) четырёхклассовая разметка, проведённая в соответствии с «Возрастной классификацией информационной продукции в России»[3] (FicRARS). Корпус TB, описанный в том числе в работе [8], состоит из текстов школьных учебников по обществознанию, разбитых на отдельные предложения. Так как корпус TB достаточно мал (менее 50000 предложений), мы объединили тексты из нескольких классов в категории:

[2]Так, средняя длина текстов в Газетном корпусе НКРЯ составляет 20 предложений.

[3]Введена согласно Федеральному закону РФ №436-ФЗ от 2010-12-23 «О защите детей от информации, причиняющей вред их здоровью и развитию».

Table 1. The characteristics of the datasets

Таблица 1. Характеристика использованных корпусов текстов

Характеристика	Fic		TB	RL	BR
Всего фрагментов	58184		449	9230	5795
Всего словоупотреблений	26252666		335436	4888290	2897003
Всего различных слов	304731		18661	103875	55577
Средняя длина текста (в словах)	918.64		747.07	1053.28	984.75
Средняя длина предложения (в словах)	13.12		10.67	14.95	13.92
Количество классов	2	4	3	3	5
Минимальный размер класса (в фрагментах)	28765	7014	98	1499	805
Максимальный размер класса (в фрагментах)	29419	21124	200	5390	1428

5–7, 8–9 и 10–11 классы. Корпусы RL и BR были представлены в работе [6]. Первый из них состоит из текстов произведений, рекомендованных Министерством просвещения РФ для внеклассного чтения, второй — из произведений, которые респонденты школьного возраста чаще всего называли последними из прочитанных. Таким образом, для рассмотрения были выбраны корпусы с различной парадигмой разметки предполагаемого возраста читателя с текстами из различных доменов. В ходе предобработки тексты каждого из корпусов Fic, RL и BR были случайным образом разделены на обучающую и тестовую выборки в соотношении 4:1, а затем разделены на фрагменты длиной 70 предложений аналогично [6]. Корпус TB был разделён на обучающую и тестовую выборки по сериям учебников: учебники за авторством А. Ф. Никитина составили обучающую выборку, за авторством Л. Н. Боголюбова — тестовую. Помимо этого, предложения были объединены в группы по 70 предложений, так как сложность отдельных предложений может значительно отличаться от сложности текста в целом. Краткая характеристика корпусов приведена в таблице 1.

Для сравнения влияния различных лингвистических характеристик мы использовали четыре алгоритма машинного обучения, упомянутые в работе [6]: метод случайного леса, метод опорных векторов, свёрточную нейронную сеть и многослойный персептрон. Для каждой из групп характеристик мы оценили качество а) алгоритмов, обученных только на вычисленных характеристиках (только для методов случайного леса и опорных векторов), б) алгоритмов, обученных только на векторном представлении текстов, и в) алгоритмов, обученных на совокупности характеристик и векторного представления. Как и в работе [6], мы рассматривали задачу классификации. Модели были спроектированы и реализованы следующим образом:

1. **Метод случайного леса (RF).** Мы использовали реализацию алгоритма из библиотеки scikit-learn [16]. Ансамбль состоял из 100 решающих деревьев, для оценки качества разбиения использовался критерий Джини. Для формирования векторного представления текста использовался алгоритма мешка слов с максимальным размером словаря 10000. В постановке с обучением на совокупности векторное представление и значения характеристик конкатенировались перед подачей в модель.
2. **Метод опорных векторов (LSVC).** Аналогично методу случайного леса мы использовали реализацию алгоритма из библиотеки scikit-learn. Для обучения использовался штраф l2 и функция потерь squared hinge. Способ формирования векторного представления текста и способ агрегации данных совпадают с таковыми для метода случайного леса.
3. **Свёрточная нейронная сеть (CNN).** Мы использовали два свёрточных слоя (256 фильтров, размер ядра 2), после каждого из которых расположен слой пулинга. Для формирования векторного представления использовалась модель geowac_lemmas_none_fasttextskipgram_300_5_2020[4] [17]. Для реализации использовался фреймворк Keras[5]. Обучение проводилось с использованием метода ранней остановки со следующими параметрами: максимальное чис-

[4] https://rusvectores.org/ru/models/
[5] https://github.com/fchollet/keras

ло эпох обучения — 100, максимальное количество эпох без улучшения результата (patience) — 20. Мы использовали оптимизатор Adam [18].

4. **Многослойный перцептрон (MLP).** Использованная архитектура состоит из трёх последовательных полносвязных слоёв из 1024 нейронов с функцией активации tanh, слоя прореживания с коэффициентом 0.5 и полносвязного слоя классификации с функцией активации softmax. Для формирования векторного представления использовалась модель distiluse-base-multilingual-cased [19] из семейства Sentence Transformers [20]. Как и в случае CNN, мы использовали фреймворк Keras, оптимизитор Adam и те же параметры обучения. При обучении с добавлением лингвистических характеристик векторное представление и значения характеристик конкатенировались перед подачей в модель.

Для предобработки текстов использовалась модель токенизации Stanza [21], обученная на данных Национального корпуса русского языка[6], для лемматизации текста использовалась библиотека pymorphy2 [22], кроме того, перед векторизацией тексты очищались от стоп-слов, входящих в набор таковых для русского языка в библиотеке NLTK [23].

Для оценки качества работы алгоритмов мы использовали метрику F1, для экспериментов с тремя и более классами использовалось взвешенное среднее, так как размеры классов значительно различаются. Запуск обучения с различными случайными зёрнами показал значительную дисперсию при зафиксированном разбиении на обучающую и тестовую выборки, в связи с чем было принято решение провести для каждой постановки пять запусков с различными случайными зёрнами и усреднить полученный результат.

2. Расширение признакового описания

2.1. Тематическое моделирование и ключевые слова

В рамках исследования признаков, связанных с тематикой текста, мы сравнили влияние на качество модели информации о тематике, полученной из латентного размещения Дирихле (LDA), и информации о ключевых словах. Этот эксперимент проводился только с текстами из корпусов Fic, RL и BR. Мы воспользовались реализацией алгоритма LDA из библиотеки scikit-learn, для каждого корпуса количество возможных тематик было равно 100.

Недавнее сравнение различных алгоритмов генерации [24] показало, что лучшего качества сразу по нескольким метрикам удаётся добиться при помощи дообучения многоязычной модели mT5 [25] на базе архитектуры T5 [26], в свою очередь представляющей собой развитие архитектуры Transformer [27]. Однако использование моделей, обучаемых с учителем, в рамках нашей задачи невозможно, так как нам не удалось найти подходящих по домену корпусов с разметкой ключевых слов. В то же время, результаты работы [24] показывают, что алгоритм RuTermExtract[7] позволяет добиться результатов, лишь немногим уступающих mT5, а по двум метрикам — и превосходящих. Поэтому мы решили использовать его в качестве алгоритма генерации ключевых. Для каждого из текстов, использованных в экспериментах, за исключением корпуса TB, при помощи RuTermExtract были предрасчитаны ключевые слова (не более 15 для каждого текста). Таким образом, для корпуса Fic было сгенерировано в совокупности 145744 ключевых слова/словосочетания, для корпуса RL — 27511, для корпуса BR — 17314. Для векторизации в случае метода случайного леса и метода опорных векторов использовался тот же алгоритм, что и для векторизации всего текста, в случае свёрточной нейронной сети и многослойного перцептрона — модель distiluse-base-multilingual-cased.

[6] https://ruscorpora.ru/

[7] https://github.com/igor-shevchenko/rutermextract

Table 2. Comparison of the performance of morpheme segmentation algorithms on the Morphodict dataset. CNN — ensemble of convolutional neural networks, CatBoost — gradient boosting algorithm over decision trees, LSTM — recurrent neural network with long short-term memory.

Таблица 2. Сравнение качества работы алгоритмов морфемной сегментации на словаре Morphodict. CNN — ансамбль свёрточных нейронных сетей, CatBoost — алгоритм градиентного бустинга над решающими деревьями, LSTM — рекуррентная нейронная сеть с долгой краткосрочной памятью.

Модель	CNN	CatBoost	LSTM
F1	0.9866	0.9326	0.9815
Precision	0.9858	0.9188	0.9800
Recall	0.9874	0.9469	0.9830
Accuracy	0.9740	0.8884	0.9661
WordAccuracy	0.9082	0.6443	0.8802

2.2. Лексические признаки

В рамках эксперимента с характеристиками текста, ассоциированными со сложностью лексики, мы сравнили две группы признаков, упоминаемых в литературе (признаки, связанные с частотностью и со словообразовательными паттернами), и признаки, основанные на анализе морфемных разборов слов. Для вычисления признаков, связанных с частотностью, использовался Частотный словарь современного русского языка: на материалах Национального корпуса русского языка [28], список словообразовательных паттернов совпадает с таковым из работы [9].

Трудность использования результатов морфемного анализа заключается в необходимости вычисления морфемных разборов слов при помощи отдельного алгоритма порождения разборов или же морфемного словаря, претендующего на полноту. При этом важно отметить, что задача построения морфемных разборов с лингвистической точки зрения трудна и, вероятно, не имеет решения [29]. Это связано с недостаточной формализацией задачи и существованием нескольких противоречащих друг другу парадигм выделения морфем.

В то же время, для решения прикладных задач, подобных исследуемой в настоящей работе, могут быть применены приближённые решения. Среди работ, посвящённых порождению морфемных разборов с разметкой типов морфем, лучшего качества удалось добиться алгоритмам машинного обучения. Так, в работе [30] представлен алгоритм генерации разборов на основе ансамбля свёрточных нейронных сетей, в работах [31, 32] проведено сравнение этого алгоритма с алгоритмом градиентного бустинга над решающими деревьями CatBoost и рекуррентной нейронной сетью с долгой краткосрочной памятью соответственно. Проанализировав использовавшийся в этих работах для обучения и тестирования словарь морфемных разборов на базе Словообразовательного словаря русского языка [33], мы обнаружили большую долю некорректно размеченных разборов [34]. В связи с этим, мы провели повторное сравнение моделей на материале словаря морфемных разборов Morphodict, использующегося в Национальном корпусе русского языка[8] (разработан на базе Словаря морфем русского языка [35]). Этот словарь содержит разборы для 75649 лемм. Всего в словаре встречается 8079 различных морфем, из которых 7148 имеют тип «корень». В среднем на слово приходится 4.12 морфем, а каждая морфема встречается в словаре 38.56 раз.

Для сравнения моделей мы провели кросс-валидацию на пяти непересекающихся выборках. Качество работы алгоритма оценивалось по метрикам, предложенным в работе [30]: F1, Precision, Recall — F-мера, точность и полнота для границ морфем без учёта их типа, Accuracy — доля верно выделенных (с учётом типа) морфем, WordAccuracy — доля полностью верных разборов. Результаты сравнения приведены в таблице 2.

[8] https://ruscorpora.ru/

Лучший результат согласно всем пяти метрикам показал алгоритм на основе ансамбля свёрточных нейронных сетей. По результатам эксперимента было решено использовать гибридное решение: для лемм, присутствующих в словаре, разбор берётся из словаря, в противном случае — порождается ансамблем свёрточных нейронных сетей, обученным на полном словаре. Итоговыми вычисляемыми характеристиками, сформировавшими набор морфемных признаков, стали (среднее, максимальное и медиана вычислялись относительно текста целиком):

1. доля уникальных морфем;
2. доля уникальных корней;
3. среднее/максимальное/медианное число морфем в слове;
4. среднее/максимальное/медианное число корней в слове;
5. доля слов, содержащих соединительные гласные;
6. средняя/максимальная/медианная суммарная длина суффиксов в слове;
7. доля слов, содержащих хотя бы один из суффиксов *-ни-*, *-ени-*, *-ц-*, *-ость-*, *-ств-*, *-ящ-*, *-ющ-*, *-еск-*, *-ист-*, *-изм-*, *-лищ-*, *-атор-*, *-тор-*, *-тель-*, *-ирова-*, *-иров-*[9].

2.3. Синтаксические признаки

Для описания синтаксических деревьев мы воспользовались форматом, описанным в рамках проекта UniversalDependencies[10]. Для порождения деревьев использовалась модель Stanza [21], обученная на материале корпуса SynTagRus[11]. В качестве базового набора синтаксических признаков Syn_{old} мы использовали набор признаков, описанный в работе [6]. Его мы сравнили с созданными наборами признаков Syn_{rich} и Syn_{best}.

Для создания набора Syn_{rich} мы выделили 46 синтаксических признаков, 36 из которых соответствуют количеству рёбер соответствующего типа в синтаксическом дереве предложения. Помимо этого для каждого предложения в качестве признаков вычислялись:

- глубина дерева;
- количество вершин;
- количество внутренних вершин;
- количество листьев;
- среднее линейное расстояние между вершиной и её непосредственным предком;
- максимальное линейное расстояние между вершиной и её непосредственным предком;
- количество вершин ровно с одним потомком;
- количество вершин ровно с двумя потомками;
- количество вершин ровно с тремя потомками;
- количество вершин ровно с четырьмя и более потомками.

На основании полученных для каждого предложения значений для текста в целом вычислялись четыре характеристики: среднее, медианное и максимальное значение признака в тексте, а также стандартное отклонение. Полученные 184 признака составили набор Syn_{rich}.

Далее на основании полученного набора Syn_{rich} мы построили набор Syn_{best}. Для этого мы оценили значимость каждого из признаков, входящих в Syn_{rich} при помощи нескольких подходов. В качестве основного метода оценки мы использовали взаимную информацию [36] характеристики и метки текста. Эксперименты проводились на материале корпусов TB и BR. Для каждого из корпусов были найдены пять характеристик с наибольшим значением этой величины. Для корпуса TB такими характеристиками оказались (в скобках указано значение взаимной информации):

- среднее количество рёбер с типом nmod (1.280);

[9] Список суффиксов был определён на основании работы [9] и консультаций с экспертами в области преподавания русского языка.

[10] https://universaldependencies.org/

[11] https://huggingface.co/stanfordnlp/stanza-ru/blob/main/models/depparse/syntagrus_charlm.pt

- средняя глубина дерева (1.068);
- среднее число внутренних вершин (0.956);
- среднее количество рёбер с типом amod (0.951);
- среднее число вершин, имеющих ровно двух потомков (0.825);

для корпуса BR:

- средняя глубина дерева (0.727);
- среднее количество рёбер с типом nmod (0.727);
- среднее количество рёбер с типом det (0.713);
- среднее число внутренних вершин (0.707);
- среднее число вершин, имеющих ровно двух потомков (0.825).

В совокупности таким образом в набор Syn_{best} было добавлено шесть признаков. Далее этот список был расширен при помощи серии экспериментов. В ходе каждого из них мы обучали модель регрессии, оценивали значимость признаков при помощи трёх метрик: Mean Decrease in Impurity [37], Permutation importance [38] и Drop-column importance (эта метрика вычисляется схоже с Permutation importance, однако вместо перемешивания значений в столбце он просто удаляется, а затем модель переучивается без него). Затем для каждой из метрик определяли пять наиболее значимых признаков. Всего было использовано три алгоритма обучения: случайный лес, алгоритм градиентного бустинга над решающими деревьями и метод опорных векторов (для этого алгоритма замерялись только метрики Permutation importance и Drop-column importance). Таким образом, мы собрали 16 списков и дополнили набор Syn_{best} теми признаками, которые оказались хотя бы в двух из них. Тем самым мы расширили набор до 21 признака:

- среднее число вершин, имеющих ровно двух потомков;
- среднее число внутренних вершин;
- среднее число вершин в дереве;
- средняя глубина дерева;
- среднее количество рёбер с типом amod;
- среднее количество рёбер с типом compound;
- среднее количество рёбер с типом conj;
- среднее количество рёбер с типом det;
- среднее количество рёбер с типом nmod;
- среднее количество рёбер с типом nsubj;
- максимальное число листьев;
- максимальное число внутренних вершин;
- максимальное число вершин в дереве;
- медианное количество рёбер с типом nmod;
- медианное число внутренних вершин;
- медианное число вершин в дереве;
- стандартное отклонение числа листьев;
- стандартное отклонение числа вершин в дереве;
- стандартное отклонение числа рёбер с типом conj;
- стандартное отклонение числа рёбер с типом det;
- стандартное отклонение среднего линейного расстояния между вершиной и её непосредственным предком.

3. Результаты

Результаты проведённых экспериментов перечислены в таблицах 3, 4 и 5. В совокупности можно говорить о значительных различиях в результатах в зависимости от конкретной модели и корпуса.

Table 3. The quality of the models trained using text representation, LDA topics (*Topic*), and keywords (*Keywords*). *BoW*, *FT*, *BERT* – text representation using a bag-of-words, FastText and BERT, respectively. In this and the following tables, for each dataset-model pair, the best achieved result and the results that differ from it by no more than 1 percentage point are highlighted in gray.

Таблица 3. Сравнение качества моделей, обученных с использованием векторного представления текста, тематик LDA (*Topic*) и ключевых слов (*Keywords*). *BoW*, *FT*, *BERT* — векторное представление текста при помощи мешка слов, FastText и BERT. В этой и следующих таблицах для каждой пары корпус-семейство моделей серым выделен лучший достигнутый результат и результаты, отличающиеся от него не более чем на 1 п. п.

Набор признаков	**FicRARS**	**FicChAd**	**RL**	**BR**
Метод случайного леса				
BoW	0.427 ± 0.004	0.743 ± 0.002	0.485 ± 0.010	0.298 ± 0.009
Topic	**0.450 ± 0.002**	**0.768 ± 0.003**	**0.538 ± 0.014**	0.311 ± 0.008
Keywords	0.422 ± 0.002	**0.768 ± 0.002**	0.448 ± 0.003	0.271 ± 0.004
BoW+Topic	0.437 ± 0.003	**0.764 ± 0.002**	0.489 ± 0.009	**0.338 ± 0.010**
BoW+Keywords	0.435 ± 0.003	**0.765 ± 0.001**	0.463 ± 0.009	0.309 ± 0.003
Метод опорных векторов				
BoW	0.424	0.751	0.618	0.322
Topic	**0.440**	**0.766**	**0.658**	0.308
Keywords	0.423	**0.763**	0.519	0.251
BoW+Topic	0.428	0.754	0.631	**0.333**
BoW+Keywords	0.420	**0.757**	0.645	**0.327**
Свёрточная нейронная сеть				
FT	**0.452 ± 0.011**	**0.792 ± 0.005**	0.531 ± 0.028	0.409 ± 0.029
FT+Topic	**0.460 ± 0.012**	**0.793 ± 0.003**	**0.552 ± 0.038**	**0.421 ± 0.006**
FT+Keywords	**0.461 ± 0.010**	**0.793 ± 0.007**	0.517 ± 0.040	**0.411 ± 0.024**
Многослойный перцептрон				
BERT	0.440 ± 0.032	0.666 ± 0.004	0.572 ± 0.040	0.333 ± 0.015
BERT+Topic	**0.541 ± 0.042**	**0.757 ± 0.025**	**0.594 ± 0.044**	0.351 ± 0.018
BERT+Keywords	0.467 ± 0.013	0.710 ± 0.008	0.541 ± 0.042	**0.374 ± 0.005**

Ни в одной из трёх серий экспериментов нам не удалось обнаружить полного превосходства одной группы признаков над другой.

Эксперименты с тематическими маркерами LDA и ключевыми словами в большинстве случаев показали превосходство первых. В постановке с обучением только на вычисленных характеристиках среднее качество метода случайного леса и метода опорных векторов, обученных только на маркерах LDA, превзошло хотя бы на 1 п. п. качество аналогичных моделей, обученных только на ключевых, в 6 экспериментах из 8. Более того, в 7 из 8 экспериментов эти модели превзошли и модели, обученные на векторном представлении текста против 2 из 8 для ключевых. В случае обучения на совокупности векторного представления и лингвистических характеристик ситуация менее выраженная: из 16 проведённых экспериментов превосходство хотя бы на 1 п. п. в 7 случаях было достигнуто моделями, обученными с добавлением маркеров LDA, в 2 — с ключевыми, в остальных 7 случаях разница составила менее 1 п. п.

В отличие от эксперимента с тематическими признаками, модели, обученные только на лексических признаках, почти всегда показывали качество хуже, чем аналогичные модели, обученные на векторном представлении текста. Только в одном случае (набор *Morpheme*, метод случайного леса, корпус RL) модель продемонстрировала значительное превосходство над базовой (почти на 10 п. п.). При этом в большинстве случаев использование любого из трёх наборов в совокупности с векторным

Table 4. The quality of the models trained using text representation, information about frequency (*Freq*), occurrence of word-formation patterns (*Pattern*) and features of morphemic word parsing (*Morpheme*). *BoW*, *FT*, *BERT* — text representation using a bag-of-words, FastText and BERT, respectively.

Таблица 4. Сравнение качества моделей, обученных с использованием векторного представления текста, сведений о частотности (*Freq*), встречаемости словообразовательных паттернов (*Pattern*) и особенностях морфемных разборов слов (*Morpheme*). *BoW*, *FT*, *BERT* — векторное представление текста при помощи мешка слов, FastText и BERT, соответственно.

Набор признаков	FicRARS	FicChAd	TB	RL	BR
Метод случайного леса					
BoW	**0.427 ± 0.004**	**0.743 ± 0.002**	**0.722 ± 0.021**	0.485 ± 0.010	0.298 ± 0.009
Freq	0.369 ± 0.001	0.669 ± 0.002	0.680 ± 0.012	0.433 ± 0.003	0.265 ± 0.012
Pattern	0.345 ± 0.002	0.616 ± 0.001	0.606 ± 0.017	0.462 ± 0.006	0.248 ± 0.006
Morpheme	0.380 ± 0.002	0.666 ± 0.003	0.680 ± 0.009	**0.582 ± 0.010**	0.239 ± 0.004
BoW+Freq	**0.424 ± 0.003**	**0.746 ± 0.002**	0.711 ± 0.033	0.475 ± 0.007	0.286 ± 0.009
BoW+Pattern	**0.425 ± 0.002**	**0.747 ± 0.003**	**0.728 ± 0.026**	0.485 ± 0.013	0.301 ± 0.003
BoW+Morpheme	**0.431 ± 0.003**	**0.752 ± 0.002**	**0.718 ± 0.030**	0.484 ± 0.005	**0.313 ± 0.003**
Метод опорных векторов					
BoW	**0.424**	**0.751**	**0.813**	**0.618**	**0.322**
Freq	0.355	0.654	0.640	0.449	0.286
Pattern	0.351	0.620	0.633	0.421	0.225
Morpheme	0.376	0.691	0.662	0.467	0.214
BoW+Freq	**0.424**	**0.750**	**0.813**	**0.622**	**0.317**
BoW+Pattern	**0.426**	**0.752**	**0.813**	**0.623**	**0.328**
BoW+Morpheme	**0.426**	**0.756**	**0.807**	**0.628**	**0.325**
Свёрточная нейронная сеть					
FT	0.452 ± 0.011	**0.792 ± 0.005**	0.538 ± 0.037	0.531 ± 0.028	0.409 ± 0.029
FT+Freq	0.454 ± 0.020	**0.796 ± 0.003**	0.560 ± 0.026	**0.563 ± 0.022**	**0.424 ± 0.014**
FT+Pattern	**0.474 ± 0.037**	**0.794 ± 0.004**	0.550 ± 0.053	0.533 ± 0.021	**0.422 ± 0.031**
FT+Morpheme	0.455 ± 0.011	**0.793 ± 0.002**	**0.615 ± 0.068**	**0.573 ± 0.035**	0.413 ± 0.010
Многослойный перцептрон					
BERT	0.440 ± 0.032	0.666 ± 0.004	0.513 ± 0.030	0.572 ± 0.040	0.333 ± 0.015
BERT+Freq	**0.477 ± 0.016**	**0.697 ± 0.005**	0.659 ± 0.021	0.528 ± 0.031	0.323 ± 0.015
BERT+Pattern	0.441 ± 0.023	0.677 ± 0.008	0.629 ± 0.010	0.563 ± 0.074	0.340 ± 0.005
BERT+Morpheme	**0.478 ± 0.020**	**0.690 ± 0.010**	**0.719 ± 0.007**	**0.611 ± 0.035**	**0.358 ± 0.014**

представлением для метода случайного леса и метода опорных векторов не давало значительного изменения качества. Улучшение хотя бы на 1 п. п. было достигнуто только один раз. Иначе обстоят дела для нейросетевых моделей. Для всех пар корпус-модель, кроме пары FicChAd+CNN, использование дополнительной информации позволило добиться прироста качества на 1 п. п. хотя бы для одного из наборов признаков. Попарное сравнение трёх наборов показывает значительную неоднородность, однако в 5 из 10 случаев лучший результат с отрывом более, чем на 1 п. п., достигнут с использованием набора признаков, связанных с особенностями морфемного строения, а ещё в трёх случаях эта модель оказалась в числе лучших (разница менее 1 п. п.).

Для метода случайного леса и метода опорных векторов, обученных с использованием синтаксических признаков, ситуация схожа с аналогичными экспериментами с лексическими признаками. В большинстве ситуаций базовая модель, обученная на векторном представлении текста, либо превосходит модели с добавлением информации, либо лишь незначительно уступает им. Обратная ситуация наблюдается только в трёх парах корпус-модель: TB+RF (лучшая модель использует

Table 5. The quality of the models trained using text representation and the syntactic feature sets Syn_{old}, Syn_{rich}, and Syn_{best} defined in subsection 2.3. *BoW*, *FT*, *BERT* — text representation using a bag-of-words, FastText and BERT, respectively.

Таблица 5. Сравнение качества моделей, обученных с использованием векторного представления текста и наборов синтаксических признаков Syn_{old}, Syn_{rich} и Syn_{best}, определённых в подразделе 2.3. *BoW*, *FT*, *BERT* — векторное представление текста при помощи мешка слов, FastText и BERT, соответственно.

Набор признаков	FicRARS	FicChAd	TB	RL	BR
Метод случайного леса					
BoW	**0.427 ± 0.004**	**0.743 ± 0.002**	0.722 ± 0.021	0.485 ± 0.010	**0.298 ± 0.009**
Syn_{old}	0.375 ± 0.001	0.675 ± 0.001	0.680 ± 0.007	**0.605 ± 0.007**	0.256 ± 0.007
Syn_{rich}	0.386 ± 0.003	0.695 ± 0.001	0.703 ± 0.007	0.573 ± 0.004	0.286 ± 0.010
Syn_{best}	0.366 ± 0.003	0.670 ± 0.002	0.687 ± 0.005	0.566 ± 0.003	0.271 ± 0.007
BoW + Syn_{old}	**0.432 ± 0.001**	0.731 ± 0.004	0.720 ± 0.015	0.485 ± 0.008	**0.292 ± 0.010**
BoW + Syn_{rich}	0.415 ± 0.004	0.731 ± 0.004	0.707 ± 0.006	0.483 ± 0.010	0.262 ± 0.008
BoW + Syn_{best}	**0.427 ± 0.004**	**0.738 ± 0.002**	**0.736 ± 0.026**	0.484 ± 0.005	0.262 ± 0.005
Метод опорных векторов					
BoW	**0.424**	**0.751**	**0.813**	0.618	**0.322**
Syn_{old}	0.373	0.680	0.704	0.606	0.227
Syn_{rich}	0.394	0.714	0.655	0.572	0.306
Syn_{best}	**0.424**	0.686	0.661	0.579	0.266
BoW + Syn_{old}	**0.425**	**0.752**	**0.804**	**0.629**	**0.323**
BoW + Syn_{rich}	**0.421**	**0.752**	**0.807**	**0.627**	**0.315**
BoW + Syn_{best}	**0.424**	**0.752**	0.800	**0.621**	**0.317**
Свёрточная нейронная сеть					
FT	0.452 ± 0.011	**0.792 ± 0.005**	0.538 ± 0.037	0.531 ± 0.028	0.409 ± 0.029
FT + Syn_{old}	**0.464 ± 0.023**	**0.793 ± 0.004**	0.598 ± 0.112	0.569 ± 0.016	**0.425 ± 0.018**
FT + Syn_{rich}	**0.454 ± 0.011**	**0.793 ± 0.003**	**0.684 ± 0.050**	**0.584 ± 0.045**	0.416 ± 0.022
FT + Syn_{best}	0.448 ± 0.012	**0.793 ± 0.003**	**0.680 ± 0.078**	0.552 ± 0.044	**0.427 ± 0.032**
Многослойный перцептрон					
BERT	0.440 ± 0.032	0.666 ± 0.004	0.513 ± 0.030	0.572 ± 0.040	**0.333 ± 0.015**
BERT + Syn_{old}	0.481 ± 0.018	**0.704 ± 0.003**	0.727 ± 0.010	**0.615 ± 0.050**	**0.337 ± 0.017**
BERT + Syn_{rich}	**0.497 ± 0.013**	**0.714 ± 0.009**	**0.771 ± 0.012**	**0.625 ± 0.035**	0.315 ± 0.017
BERT + Syn_{best}	0.464 ± 0.018	0.701 ± 0.012	0.741 ± 0.007	**0.618 ± 0.020**	**0.336 ± 0.017**

комбинацию векторного представления и набора признаков Syn_{best}), RL+RF (лучшая модель использует только набор признаков Syn_{old}) и RL+LSVC (в этом случае лучший результат показала модель, использующая комбинацию векторного представления и набора признаков Syn_{old}, а модели с использованием Syn_{rich} и Syn_{best} отстали менее, чем на 1 п. п.). При этом модели, обученные на комбинации векторного представления и набора признаков Syn_{best} ни разу не отстали более, чем на 1 п. п., от моделей на комбинации векторного представления и набора признаков Syn_{rich}, а в двух случаях (с методом случайного леса) обошли их более, чем на 1 п. п.. В случае нейросетевых алгоритмов в большинстве пар корпус-модель лучший или один из лучших результатов показали модели, обученные с использованием набора признаков Syn_{rich} В особенности это заметно для многослойного перцептрона, где лучший результат достигнут при помощи этих моделей в 4 из 5 случаев. Попарное сравнение наборов Syn_{old} и Syn_{best} не позволяет выделить превосходящий: Syn_{old} лучше в трёх случаях, Syn_{best} — в двух, в остальных пяти разница составляет менее 1 п. п.

Заключение

В данной работе рассматривается задача формирования и расширения признакового описания при оценке сложности текста. Мы провели три серии экспериментов, в ходе которых сравнили новые наборы тематических, лексических и синтаксических признаков с ранее описанными. В качестве тематических признаков были изучены автоматически генерируемые ключевые слова, в качестве лексических — признаки, связанные с морфемным строением слов, в качестве синтаксических — обширный набор признаков, включающий информацию о типах рёбер, разветвлённости и глубине деревьев. Мы провели сравнение с использованием четырёх различных алгоритмов машинного обучения на материале четырёх корпусов (в одном из которых использовано два типа возрастных меток), что позволило получить более обширное представление о применимости признаков.

Проведённый анализ показал, что в большинстве ситуаций ключевые слова показали худший результат в сравнении с тематическими маркерами, полученными при помощи латентного размещения Дирихле. Это может быть связано как с недостаточным качеством генерации ключевых, так и с малой применимостью ключевых слов для домена художественных текстов. Для установления точных причин необходимо проведение дополнительных исследований. В то же время использование двух других новых наборов признаков (набора признаков, связанных с морфемным разнообразием, и обширного набора синтаксических признаков) позволило улучшить качество работы в сравнении с ранее изученными способами формирования признакового описания. При этом формирование этих наборов алгоритмически сложнее (в особенности это касается набора признаков, связанных с морфемным разнообразием, где для формирования признаков текст предварительно токенизируется, лемматизируется, а затем каждой лемме сопоставляются морфемные разборы, в том числе, автоматически генерируемые при помощи свёрточной нейронной сети).

В дальнейшем мы планируем сконцентрироваться на поиске других лингвистически мотивированных признаков и наиболее эффективных комбинаций из числа ранее изученных признаков, в частности, планируется провести более подробное исследование синтаксических признаков и интерпретировать полученные результаты с точки зрения лингвистики. Кроме того, будет проведён поиск устойчивых относительно смены корпуса и типа разметки моделей. Также интересной для изучения представляется оценка алгоритмической сложности вычисления тех или иных признаков и сравнение различных моделей с точки зрения быстродействия.

References

[1] R. Flesch, “A new readability yardstick.”, *Journal of Applied Psychology*, vol. 32, no. 3, p. 221, 1948.

[2] E. Dale and J. S. Chall, “A formula for predicting readability: Instructions”, *Educational Research Bulletin*, vol. 27, pp. 37–54, 1948.

[3] R. Senter and E. A. Smith, “Automated readability index”, AMRL TR, Tech. Rep. 5302480, 1967.

[4] M. Solnyshkina, V. Ivanov, and V. Solovyev, “Readability formula for Russian texts: A modified version”, in *Proceedings of the 17th Mexican International Conference on Artificial Intelligence, Part II*, 2018, pp. 132–145. DOI: 10.1007/978-3-030-04497-8_11.

[5] A. Churunina, M. Solnyshkina, E. Gafiyatova, and A. Zaikin, “Lexical features of text complexity: The case of Russian academic texts”, *SHS Web of Conferences*, vol. 88, no. 1, p. 01 009, 2020. DOI: 10.1051/shsconf/20208801009.

[6] D. A. Morozov, A. V. Glazkova, and B. L. Iomdin, “Text complexity and linguistic features: Their correlation in English and Russian”, *Russian Journal of Linguistics*, vol. 26, no. 2, pp. 426–448, 2022. DOI: 10.22363/2687-0088-30132.

[7] N. Karpov, J. Baranova, and F. Vitugin, "Single-sentence readability prediction in Russian", in *Analysis of Images, Social Networks and Texts*, Cham: Springer International Publishing, 2014, pp. 91–100. DOI: 10.1007/978-3-319-12580-0_9.

[8] V. V. Ivanov, M. I. Solnyshkina, and V. D. Solovyev, "Efficiency of text readability features in Russian academic texts", *Komp'juternaja Lingvistika I Intellektual'nye Tehnologii*, vol. 17, pp. 267–283, 2018.

[9] O. Blinova and N. Tarasov, "A hybrid model of complexity estimation: Evidence from Russian legal texts", *Frontiers in Artificial Intelligence*, vol. 5, p. 1 008 530, 2022. DOI: 10.3389/frai.2022.1008530.

[10] U. Isaeva and A. Sorokin, "Investigating the robustness of reading difficulty models for Russian educational texts", in *Recent Trends in Analysis of Images, Social Networks and Texts*, Cham: Springer International Publishing, 2021, pp. 65–77. DOI: 10.1007/978-3-030-71214-3_6.

[11] A. N. Laposhina, T. S. Veselovskaya, M. U. Lebedeva, and O. F. Kupreshchenko, "Lexical analysis of the Russian language textbooks for primary school: Corpus study", in *Komp'juternaja Lingvistika I Intellektual'nye Tehnologii*, vol. 18, 2019, pp. 351–363.

[12] V. Solovyev, V. Ivanov, and M. Solnyshkina, "Readability formulas for three levels of Russian school textbooks", *Investigations on Applied Mathematics and Informatics. Part II–1*, vol. 529, pp. 140–156, 2023.

[13] A. N. Laposhina, M. Y. Lebedeva, and A. A. Berlin Khenis, "Word frequency and text complexity: An eye-tracking study of young Russian readers", *Russian Journal of Linguistics*, vol. 26, no. 2, pp. 493–514, 2022. DOI: 10.22363/2687-0088-30084.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation", *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[15] A. Glazkova, Y. Egorov, and M. Glazkov, "A comparative study of feature types for age-based text classification", in *Analysis of Images, Social Networks and Texts*, Cham: Springer International Publishing, 2021, pp. 120–134, ISBN: 978-3-030-72610-2. DOI: 10.1007/978-3-030-72610-2_9.

[16] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python", *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] A. Kutuzov and E. Kuzmenko, "WebVectors: A toolkit for building web interfaces for vector semantic models", in *Analysis of Images, Social Networks and Texts*, Springer, 2017, pp. 155–161. DOI: 10.1007/978-3-319-52920-2_15.

[18] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG].

[19] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365.

[20] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019. DOI: 10.18653/v1/D19-1410.

[21] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, 2020, pp. 101–108. DOI: 10.18653/v1/2020.acl-demos.14.

[22] M. Korobov, "Morphological analyzer and generator for Russian and Ukrainian languages", in *International Conference on Analysis of Images, Social Networks and Texts*, Springer, 2015, pp. 320–332. DOI: 10.1007/978-3-319-26123-2_31.

[23] E. Loper and S. Bird, "NLTK: The natural language toolkit", in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 2002, pp. 63–70.

[24] A. Glazkova, D. Morozov, M. Vorobeva, and A. Stupnikov, "Keyphrase generation for the Russian-language scientific texts using mT5", *Modeling and Analysis of Information Systems*, vol. 30, no. 4, pp. 418–428, 2023. DOI: 10.18255/1818-1015-2023-4-418-428.

[25] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41.

[26] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer", *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[27] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

[28] O. Lyashevskaya and S. Sharov, *Chastotnyj slovar' sovremennogo russkogo yazyka: na materialah Nacional'nogo korpusa russkogo yazyka*. Azbukovnik, 2009, in Russian, ISBN: 9785911720247.

[29] B. L. Iomdin, "How to define words with the same root?", *Russian Speech = Russkaya Rech'*, vol. 1, pp. 109–115, 2019, in Russian. DOI: 10.31857/S013161170003980-7.

[30] A. Sorokin and A. Kravtsova, "Deep convolutional networks for supervised morpheme segmentation of Russian language", in *Artificial Intelligence and Natural Language*, Cham: Springer International Publishing, 2018, pp. 3–10. DOI: 10.1007/978-3-030-01204-5_1.

[31] E. I. Bolshakova and A. S. Sapin, "Comparing models of morpheme analysis for Russian words based on machine learning", in *Komp'juternaja Lingvistika I Intellektual'nye Tehnologii*, vol. 18, 2019, pp. 104–113.

[32] E. Bolshakova and A. Sapin, "Bi-LSTM model for morpheme segmentation of Russian words", in *Artificial Intelligence and Natural Language*, Cham: Springer International Publishing, 2019, pp. 151–160. DOI: 10.1007/978-3-030-34518-1_11.

[33] A. N. Tikhonov, *Slovoobrazovatel'nyi slovar' russkogo yazyka*. Moscow: Russkiy yazyk, 1990, in Russian.

[34] T. Garipov, D. Morozov, and A. Glazkova, "Generalization ability of CNN-based Morpheme Segmentation", in *2023 Ivannikov Ispras Open Conference (ISPRAS)*, 2024, pp. 58–62.

[35] A. I. Kuznetsova and T. F. Efremova, *Dictionary of Morphemes of the Russian Language*. Firebird Publications, Incorporated, 1986, 1136 pp.

[36] T. Cover and A. Joy, "Entropy, relative entropy, and mutual information", in *Elements of Information Theory*. John Wiley & Sons, Ltd, 2005, ch. 2, pp. 13–55, ISBN: 9780471748823. DOI: 10.1002/047174882X.ch2.

[37] L. Breiman, J. Friedman, C. J. Stone, and R. Olshen, *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

[38] A. Altmann, L. Tolosi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure", *Bioinformatics (Oxford, England)*, vol. 26, no. 10, pp. 1340–1347, 2010. DOI: 10.1093/bioinformatics/btq134.