Московский государственный университет имени М. В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Веселов Арсений Сергеевич

Оценивание когнитивной сложности текста при помощи квантильного подхода и агрегирования

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:

д.ф.-м.н., профессор

К.В. Воронцов

Москва, 2023

Содержание

1 Введение

Для задачи оценивания сложности текста было разработано множество индексов удобочитаемости. Большинство из них использует в своей основе линейную комбинацию некоторых тривиальных статистических параметров текста, основанных на количестве букв, слогов, слов и предложений. В данной работе продолжается исследование и совершенствование обобщённого квантильного подхода к оцениванию когнитивной сложности текста на разных уровнях языка: на фонетическом, морфемном, лексическом и синтаксическом. Идея такого подхода была впервые представлена Еремеевым М.А. и Воронцовым К.В. в [1]. В его основе лежит определение токенов с аномальной частотой их оценок сложности по эмпирическим распределениям. Эмпирические распределения строятся по референтному корпусу текстов, в качестве которого используется русскоязычная Википедия. Настоящая работа посвящена исследованию агрегирования отдельных квантильных моделей, чтобы учитывать информацию с разных уровней языка, и в этом заключается её новизна. Мы обучаем агрегированные модели на выборке пар фрагментов текстов, которую создали на основе учебников по обществознанию разных учебных классов. В данной работе были проведены эксперименты по сравнению точности моделей с адаптированными индексами удобочитаемости, в том числе и между отдельными парами учебных классов. Также был проведён анализ вклада отдельных компонентов в агрегированную модель (абляционные исследования (ablation study)) и анализ зависимости точности ранжирования от средней длины фрагмента текста в датасете. Проведённые в работе эксперименты демонстрируют у предлагаемого подхода более высокую точность ранжирования текстов по когнитивной сложности в сравнении с индексами удобочитаемости. Целью проведения данного исследования является создание одного из важных компонентов системы рекомендации научно-образовательного контента.

2 Обзор индексов удобочитаемости

Индексы удобочитаемости были разработаны лингвистами для оценки сложности текста учебной литературы. Многие из них изначально разрабатывались для системы образования США, а потому были адаптированы для английского языка.

Автоматический индекс удобочитаемости (automated readability index (ARI)) был разработан Сэнтером (R.J. Senter) и Смитом (E.A. Smith) в 1967 году [2]. Данный индекс показывает приблизительный номер класса в системе образования США, после прохождения которого анализируемый текст будет понятен для учащегося в нём человека. ARI имеет следующую формулу расчёта для документа d на английском языке:

$$\mathrm{ARI}(d) = 4.71 \times \frac{C}{W} + 0.5 \times \frac{W}{S} - 21.43,$$

где C — количество букв и цифр, W — количество слов, S — количество предложений в тексте документа d.

Индекс удобочитаемости Läsbarhetsindex (LIX) был разработан шведским учёным Карлом-Хуго Бьёрнссоном (Carl-Hugo Björnsson) в 1968 году [3]. LIX не учитывает, на каком языке написан текст. С повышением сложности текста значение индекса возрастает. Формула расчёта LIX для документа d на любом языке:

$$\mathrm{LIX}(d) = \frac{A}{B} + 100 \times \frac{C}{A},$$

где A — количество слов в тексте, B — количество предложений в тексте, C — количество слов длиннее 6 букв. Чем выше значение индекса LIX, тем сложнее текст.

В 1969 году Маклафлином (G. Harry McLaughlin) разработал индекс удобочитаемости SMOG (Simple Measure of Gobbledygook) [4]. Он оценивает количество лет обучения, необходимых для понимания текста. SMOG имеет следующую формулу расчёта для документа d на английском языке:

$$\mathrm{SMOG}(d) = 1.0430\sqrt{A \times \frac{30}{B}} + 3.1291,$$

где A — количество многосложных слов в тексте (3 и более слогов для английского языка), B — количество предложений в тексте (минимум 30).

Индекс Коулман — Лиау (Coleman — Liau index (CLI)), разработанный в 1975 году Мэри Коулман (Meri Coleman) и Т.Л. Лиау (T.L. Liau) [5], выдаёт приблизительный номер класса в системе образования США, закончив который человек должен будет понимать анализируемый текст. CLI имеет следующую формулу расчёта для документа d на английском языке:

$$\mathrm{CLI}(d) = 0.0588 \times L - 0.296 \times S - 15.8,$$

где L — среднее количество букв на 100 слов, S — среднее количество предложений на 100 слов.

В 1948 году Рудольф Флеш (Rudolf Flesch) разработал индекс удобочитаемости Флеша (Flesch reading-ease score (FRES)) [6] для текстов на английском языке. С повышением сложности текста значение индекса убывает. Формула расчёта FRES:

$$\mathrm{FRES}(d) = 206.835 - 1.015 \times \mathrm{ASL} - 84.6 \times \mathrm{ASW},$$

где ASL — средняя длина предложения в словах (average sentence length), ASW — средняя длина слова в слогах (average number of syllables per word). Результат FRES интерпретируется так: чем *меньше* значение FRES, тем сложнее текст.

В 1975 году по заказу ВМС США Питером Кинкейдом (J. Peter Kincaid) была разработана формула (Flesch–Kincaid grade level (FKGL)) [7]. Этот индекс удобочитаемости выводит приблизительный номер класса в американской системе

образования. FKGL имеет следующую формулу расчёта для документа d на английском языке:

$$\text{FKGL}(d) = 0.39 \times \text{ASL} + 11.8 \times \text{ASW} - 15.59.$$

Эстонский лингвист Тулдава Ю.А. (Juhan Tuldava) в 1975 году предложил индекс удобочитаемости [8], который в данной работе будет называться Tuldava Index (TI). TI, как и индекс LIX, не учитывает, на каком языке написан текст. Формула расчёта TI для документа d на любом языке:

$$\text{TI}(d) = \text{ASW} \times lg(\text{ASL}).$$

В данной работе будет оцениваться сложность текстов на русском языке, хотя у реализованного алгоритма есть возможность работать и с английскими текстами. Поэтому предлагаемый частотный подход будет сравниваться с вышеперечисленными индексами удобочитаемости, адаптированными для русского языка.

Существенный вклад в разработку формулы читабельности для текстов на русском языке внесла Оборнева И.В., адаптировавшая в 2006 году индексы FRES и FKGL [9]:

$$\text{FRES}_{\text{ru}}(d) = 206.835 - 1.3 \times \text{ASL} - 60.1 \times \text{ASW},$$

$$\text{FKGL}_{\text{ru}}(d) = 0.5 \times \text{ASL} + 8.4 \times \text{ASW} - 15.59.$$

Для адаптации формул к тексту на русском языке Оборнева И.В. осуществила сравнительный анализ средней длины слова в русском и английском языках.

В последующем результат адаптации формул удобочитаемости для автоматизированного анализа текстов на русском языке был представлен Бегтиным И.В. в 2014 году [10], и собран в библиотеку ruTS для Python Шкариным Сергеем в 2021 году [11]. Для данной работы в дополнение к уже реализованным в библиотеке ruTS индексам был добавлен индекс удобочитаемости Тулдавы, а также в ней были исправлены неправильные коэффициенты в реализации индекса Коулман — Лиау. Приведём ниже формулы индексов ARI, SMOG и CLI, адаптированных для русского языка (величины без пояснений такие же, как и для формул для английского языка):

$$\text{ARI}_{\text{ru}}(d) = 6.26 \times \frac{C}{W} + 0.2805 \times \frac{W}{S} - 31.04.$$

$$\text{SMOG}_{\text{ru}}(d) = 1.1\sqrt{A \times \frac{64.6}{B}} + 0.05,$$

где A — количество многосложных слов в тексте (4 и более слогов для русского языка).

$$\text{CLI}_{\text{ru}}(d) = 0.055 \times L - 0.35 \times S - 20.33.$$

Оценивание сложности текста может применяться различными организациями к текстовым документам совершенно разных направленностей. Индекс FRES очень

популярен в США, к тому же он признаётся некоторыми государственными органами как стандарт для оценивания удобочитаемости документов и договоров. Индекс FKGL разрабатывался с целью составления текстов инструкций по применению оружия или технических средств, а сейчас он широко используется в образовательной сфере. Индекс удобочитаемости SMOG использовался для изучения сложности текстов инструкций к лекарствам и препаратам. В [12] описывается применение индексов удобочитаемости для анализа решений конституционного суда. Многие индексы применяются для оценивания понятности учебных пособий, предлагаемых учащимся разного возраста. Использование оценки сложности текстов может быть полезно для прогнозирования временных затрат на обработку нормативных актов, документов и учебной литературы.

3 Обобщённая модель сложности текста

Пусть d — произвольный документ, состоящий из токенов $x_1, \ldots, x_n$ из фиксированного конечного алфавита A_h, где h обозначает уровень языка: фонетический, морфемный, лексический или синтаксический. В данной работе в качестве токенов будут рассматриваться, в зависимости от уровня языка, соответственно буквы, слоги, слова и предложения (или структуры, описывающие часть речи и синтаксическую функцию слов). Предположим, что у каждого токена x_i из документа d возникает своя сложность обработки, обусловленная его контекстом или же его внутренней структурой, оценку которой мы обозначим за c_i. Также предположим, что у каждого токена $a \in A_h$ существует его привычная сложность обработки, выработанная в результате развития языка в культурно-историческом контексте. Если текущая сложность обработки токена $x_i = a$ в анализируемом тексте оказывается аномально высокой по сравнению с привычной сложностью обработки токенов того же типа, то будем считать, что токен x_i несёт в себе повышенную нагрузку для восприятия. Информация о привычной сложности токенов может быть извлечена из референтного корпуса текстов, обозначаемого K, который является большой коллекцией текстов умеренной сложности. Чтобы оценить, является ли токен $x_i \in d,\ x_i = a$ аномально сложным, требуется построить эмпирическое распределение оценок сложности $\hat{c}_j$ каждого токена $\hat{x}_j \in K$, такого что $\hat{x}_j = a$. Токен x_i считается аномально сложным, если его оценка сложности больше, чем γ-квантиль $C_\gamma(x_i)$ построенного эмпирического распределения для токена (см. рис. 1).

На рис. 1 высота i-ого столбца гистограммы показывает, сколько раз токен сложности i встретился в референтном корпусе. Зона аномально высокой сложности, вызывающей дополнительную нагрузку для восприятия, отмечена красным цветом. Зона пониженной сложности изображена зелёным цветом. Синяя зона соответствует привычной сложности токена.

Будем понимать под оценкой когнитивной сложности документа $S(d)$ сумму весов

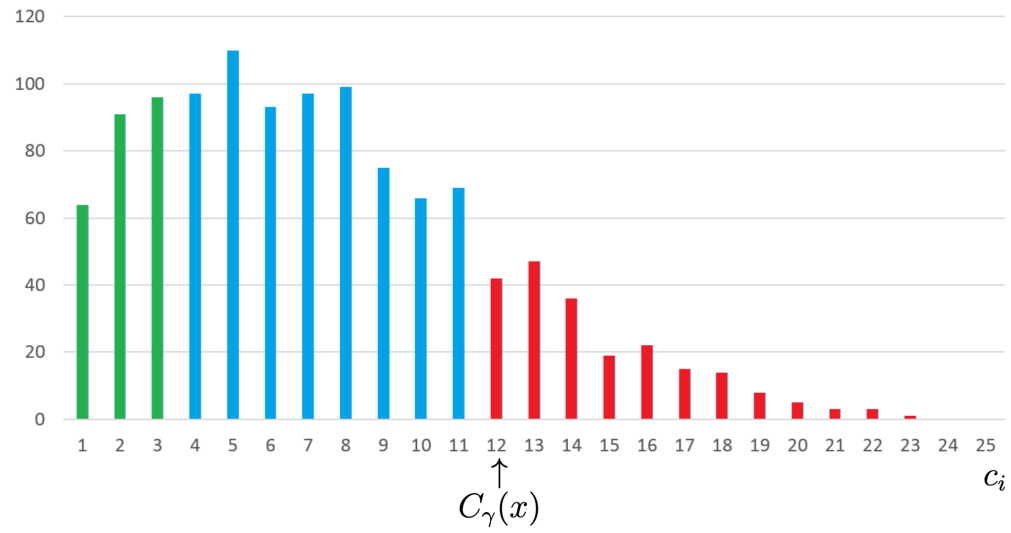

Рис. 1. Пример эмпирического распределения сложности токена

w_i токенов, имеющих аномальную сложность.

$$S(d) = \sum_{i=1}^{n} w_i \left[c_i > C_\gamma\left(x_i\right)\right],$$

где [ ] — скобка Айверсона (т.е. [истина] = 1, [ложь] = 0). Можно возвести веса в степень p и получить нелинейную сумму оценок сложности:

$$S(d) = \sum_{i=1}^{n} w_i^p \left[c_i > C_\gamma\left(x_i\right)\right], \tag{1}$$

где $p > 0$ — целое число.

Можно ввести нормированную когнитивную сложность документа d, чтобы сравнивать оценки сложности документов с разными длинами:

$$W(d) = \frac{\sum_{i=1}^{n} w_i^p \left[c_i > C_\gamma\left(x_i\right)\right]}{\sum_{i=1}^{n} w_i^p}. \tag{2}$$

Вес w_i — неотрицательная величина, монотонно неубывающая с ростом оценки сложности токена c_i, которая, в свою очередь, определена с точностью до произвольной монотонно возрастающей функции.

Для токенов, не присутствовавших в референтном корпусе, γ-квантиль берётся равной $-\infty$. В этом случае модель будет всегда считать данный токен аномально сложным.

Различные модели сложности текста можно строить, выбирая алфавит A_h и способы вычисления оценок сложности токенов c_i и весов w_i.

В таблице 1 приведено несколько примеров возможных способов задания весов.

w_i	Смысл w_i
1	число сложных токенов
$1/n \times 100\%$	процентное содержание сложных токенов
c_i	общая сложность
c_i/n	средняя сложность
$c_i - C_\gamma(x_i)$	избыточная сложность
$(c_i - C_\gamma(x_i))/n$	средняя избыточная сложность

Табл. 1. Примеры нагрузок w_i

4 Функции сложности отдельных токенов

4.1 Частотная функция

При частотном подходе к оцениванию сложности отдельных токенов будем смотреть на частоту употребления каждого типа токена в документе, иными словами, на расстояния между последовательными вхождениями одного и того же типа токена.

Обозначим за r_i расстояние между предыдущим вхождением токена x_i и его текущим вхождением в текст:

$$\ldots \boxed{x_{i-r_i} = a} \quad \underbrace{x_{i-r_i+1} \quad x_{i-r_i+2} \quad \ldots \quad x_{i-2} \quad x_{i-1} \quad \boxed{x_i = a}}_{r_i} \ldots$$

Формально:

$$r_i = \min_{1 \leqslant j < i} \{i - j \mid x_i = x_j\}.$$

В момент первого появления токена a в тексте документа d на позиции i значение r_i ещё не определено. В таком случае, r_i доопределяется так, чтобы сумма r_i для всех вхождений одного и того же токена $x_i = a$ была равна длине документа n.

Приведём пример определения значений r_i для фразы «естественный текст». Пусть A_h состоит из букв. При первом проходе возможно определить значения r_i только для девяти позиций токенов. После доопределения все значения r_i оказываются заполненными:

токен	е	с	т	е	с	т	в	е	н	н	ы	й	т	е	к	с	т
r_i исходное	–	–	–	3	3	3	–	4	–	1	–	–	7	6	–	11	4
r_i доопред.	6	5	5	3	3	3	19	4	18	1	19	19	7	6	19	11	4

Табл. 2. Пример доопределения r_i для фразы «естественный текст»

В частотной модели оценки сложности документа определим параметры c_i как некоторую убывающую функцию от r_i. В экспериментах используется следующая

функция:

$$c_i = -r_i \tag{3}$$

4.2 Сложностная функция

Рассмотрим другой способ оценивания сложности отдельного токена. Будем теперь отталкиваться от свойств внутренней структуры токена или же от свойств взаимосвязи токена с его контекстом.

В сложностном подходе к оцениванию сложности отдельных токенов будем смотреть лишь на сложности токенов, не различая их по типам. Другими словами, будем считать, что алфавит A_h состоит из единственного токена $A_h = \{a\}$.

При таком способе будет построено только одно эмпирическое распределение оценок сложности отдельных токенов общее для всех токенов из референтного корпуса текстов, а γ-квантиль не будет зависеть от типа токена: $C_\gamma(x_i) = C_\gamma$.

5 Рассматриваемые модели

В терминах, введённых при рассмотрении обобщённой модели оценивания когнитивной сложности текста, можно описать отдельные модели для разных уровней языка, задавая алфавит токенов и способ вычисления параметров c_i для каждой позиции i. Для формирования алфавитов токенов и характеристик их сложности могут использоваться готовые средства морфологического, лексического и синтаксического анализа.

5.1 Фонетический уровень

На фонетическом уровне языка в качестве токенов рассматриваются отдельные буквы (табл. 2). Подход используется частотный. Краткое название модели: $letter\_dist\_model$, где dist от слова distance, так как частотный подход основан на расстоянии между токенами.

5.2 Морфемный уровень

На морфемном уровне языка в качестве токенов можно брать либо исходные слоги, либо слоги с отсортированными в алфавитном порядке в них буквами. Таким образом, рассматриваются две частотных морфемных модели со следующими краткими названиями: $syllab\_dist\_model$, $syllabsort\_dist\_model$.

5.3 Лексический уровень

На лексическом уровне в качестве токенов рассматриваются отдельные слова. На данном уровне языка для частотной модели и для сложностной модели терминов разные словоформы одного слова будем считать эквивалентными токенами. С этой целью в качестве предобработки проводится нормализация словоформ (для русского языка — лемматизация).

Частотная модель. Основное предположение частотной лексической модели заключается в том, что чем чаще встречается одно и то же слово в анализируемом тексте, тем большую нагрузку для восприятия оно оказывает при обработке текста человеком. Краткое название модели: $lexical\_dist\_model$.

Сложностная модель длины слова. Данная модель использует длину слова как оценку его сложности. В этом случае аномально сложными являются аномально длинные слова. Краткое название модели: $lexical\_len\_model$.

Сложностная модель терминов. В рамках данной модели будем исходить из того, что узкоспециализированные термины оказывают избыточную нагрузку для восприятия при их обработке человеком. Степень этой узкоспециализированности можно оценить по частоте употребления слова в языке, для этого используется референтный корпус.

Для сложностной лексической модели терминов определим параметры c_i как убывающую функцию величины $\mathrm{count}(x_i)$, которая, в свою очередь, равна количеству вхождений слова x_i в референтный корпус. В экспериментах используется следующая функция:

$$c_i = -\mathrm{count}(x_i).$$

Краткое название модели: $lexical\_cnt\_model$.

5.4 Синтаксический уровень

На синтаксическом уровне языка токенами являются предложения или структуры, описывающие синтаксические роли и части речи слов в предложении. В данной работе для деления текста на предложения и извлечения синтаксических зависимостей и частей речи используется библиотека UDPipe [13].

Сложностная модель. Главное предположение сложностной синтаксической модели заключается в том, что аномально длинные синтаксические связи должны приводить к избыточной нагрузке для восприятия при обработке текста человеком. Данная модель берёт за оценку сложности предложения максимальную длину синтаксической связи в нём. Краткое название модели: $syntax\_len\_model$.

Частотная модель. Данная модель использует в качестве токенов структуры, описывающие часть речи и синтаксическую функцию слов. Каждая такая структура соответствует одному слову, однако само слово в дальнейшем анализе не участвует.

В рамках данной модели будем исходить из предположения, что аномально частое употребление отдельных структур приводит к повышенной нагрузке для восприятия при их обработке человеком.

Если описывать алфавит для данной модели более формально, то каждый токен $a \in A_h$ является парой $(p,\ s)$, где p принадлежит множеству всевозможных частей речи, а s — множеству всевозможных типов членов предложения. Краткое название модели: *syntaxpos_dist_model*.

6 Эксперименты

6.1 Наборы данных

В качестве референтного корпуса для экспериментов использовалась Википедия (1.5 миллиона текстов). Она была загружена в виде архива ruwiki-latest-pages-articles.xml.bz2, а затем обработана парсером WikiExtractor и после этого с дополнительной предобработкой переведена в формат, где каждой статье соответствует свой .txt документ.

В качестве набора данных для валидации результатов были использованы наборы учебников по обществознанию, подготовленные в [14]. Учебники Л.Н. Боголюбова. для 6, 7, 8, 9, 10, 10+, 11+ классов («+» обозначает версию с углублённым специализированным изучением) и учебники А.Ф. Никитина для 5, 6, 7, 8, 9, 10, 11 классов. В этом наборе данных каждый документ — это предложения из текста учебника, расположенные в случайном порядке, чтобы не нарушить авторские права. Для создания датасетов для обучения и валидации моделей тексты учебников одинаковых классов сначала объединялись, а потом нарезались на фрагменты приблизительно одинаковой длины так, чтобы в этих фрагментах содержались цельные предложения. Затем фрагменты текстов разных учебных классов объединялись в пары, где на первом месте стоит фрагмент текста из учебника с более младшим учебным классом: $D = \{(d,\ d') \mid d' \text{ сложнее, чем } d)\}$. Сложность учебника определяется по номеру класса. Номер класса учебника должен являться достаточно надёжной характеристикой, на которую можно ориентироваться при оценивании когнитивной сложности текста, поскольку учебные пособия создаются в соответствии с образовательными стандартами.

На основе набора учебников было подготовлено восемь датасетов с разным числом пар. Для этого варьировалась длина одного фрагмента текста (см. табл. 3). В каждом датасете присутствуют все возможные пары, такие что каждый фрагмент текста одного учебного класса был сопоставлен с каждым фрагментом текста каждого другого учебного класса.

Такое количество датасетов мы создаём с целью исследовать в последнем из

Название датасета	**Число пар фрагментов**	**Среднее кол-во символов в одном фрагменте**
D1	1027	94 100
D2	2532	59 850
D3	5001	42 650
D4	10 041	30 100
D5	45 058	14 200
D6	250 152	6000
D7	1 008 881	2950
D8	5 400 136	1250

Табл. 3. Датасеты

экспериментов зависимость точности ранжирования от средней длины фрагмента текста. В экспериментах же до него используются лишь датасеты D1, D2, D3, D4, потому что в них средняя длина фрагмента текста достаточно большая для того, чтобы предоставить как можно больше информации моделям и индексам удобочитаемости для оценивания сложности фрагментов. Причём ориентироваться в последующих экспериментах мы будем больше на D4, так как в этом датасете присутствует достаточно много пар фрагментов текста для того, чтобы можно было получить больше разных возможных значений критерия качества, а также обучить агрегированные модели на большем числе пар.

6.2 Критерий качества

За метрику качества взята точность (accuracy), т.е. отношение числа правильно оценённых пар фрагментов текстов к общему числу пар:

$$\text{accuracy(S)} = \frac{\sum_{(d,d')\in D} [S(d') > S(d)]}{|D|},$$

где за S обозначена модель (или индекс удобочитаемости), выдающая оценку сложности текста.

6.3 Используемые модели

В экспериментах (см. табл. 4) был произведён подбор наилучших параметров $(p,\ w_i,\ \gamma)$ (см. формулу 1) моделей, которые максимизировали критерий качества на 3-ем и 4-ом датасетах (так как в них больше пар, чем в D1 и D2, а значит потенциально возможно получить больше разных значений критерия качества) либо на аналогичных по количеству пар датасетов, построенных на серии

учебников только одного из авторов. Тип нагрузок w_i перебирался из множества $\{1,\ c_i,\ c_i/n,\ c_i - C_\gamma(x_i),\ (c_i - C_\gamma(x_i))/n\}$. Параметр γ перебирался по сетке с шагом 0.05: $\{0.01\} \cup [0.05,\ 0.1,\ 0.15,\ \ldots,\ 0.9,\ 0.95] \cup \{0.99\}$. Параметр p перебирался по сетке $[1,\ 2,\ 3,\ 4]$. Формула (2) тоже проверялась, но показала качество хуже. Кроме частотных моделей с функцией сложности (3) в экспериментах было проверено качество моделей, использующих альтернативную гипотезу о том, что чем реже одинаковые токены встречаются в тексте, тем большую нагрузку для восприятия они несут, с функцией сложности: $C_i = r_i$. Но качество у таких моделей было значительно ниже, чем у остальных, поэтому далее они представлены не будут.

№	Название модели	Гиперпараметры			Точность на датасете, %			
		w_i	p	γ	D1	D2	D3	D4
1	***letter_dist_0***	c_i/n	1	0.10	79.45	77.49	77.70	76.36
2	***letter_dist_1***	c_i/n	1	0.85	81.60	77.13	76.42	75.74
3	*letter_dist_2*	c_i	1	0.05	80.92	72.08	80.66	67.29
4	*syllab_dist_0*	c_i/n	1	0.01	63.49	76.46	78.08	78.23
5	*syllab_dist_1*	c_i	1	0.65	73.61	63.19	72.27	59.57
6	*syllabsort_dist_0*	c_i	1	0.05	79.07	67.65	82.04	76.25
7	*lexical_dist_0*	$(c_i - C_\gamma(x_i))/n$	1	0.01	75.17	76.11	84.88	82.01
8	*lexical_dist_1*	c_i	1	0.99	82.38	76.94	85.64	76.17
9	***lexical_len_0***	$(c_i - C_\gamma(x_i))/n$	1	0.55	92.02	90.72	89.58	89.57
10	***lexical_len_1***	c_i/n	1	0.85	88.41	87.52	87.00	86.86
11	***lexical_len_2***	c_i/n	1	0.45	92.11	90.84	91.10	91.08
12	***lexical_len_3***	$(c_i - C_\gamma(x_i))/n$	1	0.30	**93.48**	**92.06**	91.70	**91.28**
13	***lexical_len_4***	c_i/n	2	0.65	90.36	89.38	**92.76**	87.72
14	*lexical_cnt_0*	c_i/n	2	0.35	70.79	61.77	72.02	63.10
15	*lexical_cnt_1*	c_i/n	2	0.15	84.32	83.10	87.16	80.78
16	*lexical_cnt_2*	c_i	1	0.85	63.68	57.70	63.25	57.50
17	*lexical_cnt_3*	c_i	1	0.45	73.81	67.69	71.25	60.92
18	***syntax_len_0***	c_i/n	2	0.01	88.61	83.77	86.14	83.95
19	***syntax_len_1***	c_i/n	2	0.35	88.51	83.81	85.80	83.89
20	*syntaxpos_dist_0*	c_i	1	0.45	81.60	81.67	85.58	78.99
21	*syntaxpos_dist_1*	c_i	1	0.35	83.93	82.39	86.30	80.84

Табл. 4. Подобранные параметры и точность отдельных моделей. Жирными горизонтальными линиями отделены разные типы моделей. Модели, имена которых выделены жирным, в абляционных исследованиях показали наибольший вклад в ансамбль

В итоге была отобрана 21 модель, которая будет использоваться в экспериментах по агрегации. Из отдельных моделей наилучшее качество показали лексические сложностные модели длины слова, превзойдя все индексы удобочитаемости, точность которых на этих же датасетах представлена в таблице 5. Из индексов удобочитаемости наилучшую точность демонстрируют FKGL_{ru} и FRES_{ru}. Если ориентироваться только на датасет D4, то лучшим индексом является FRES_{ru}.

Индекс	**Точность на датасете, %**			
	D1	**D2**	**D3**	**D4**
FKGL_{ru}	**91.04**	**90.00**	89.94	89.49
FRES_{ru}	90.75	**90.00**	**90.30**	**90.50**
CLI_{ru}	89.97	89.26	89.76	89.09
SMOG_{ru}	90.26	88.63	88.24	87.80
ARI_{ru}	90.36	89.69	90.14	89.64
LIX	90.65	89.22	89.44	88.79
TI	90.94	89.97	89.92	89.55

Табл. 5. Результаты индексов удобочитаемости на D1, D2, D3 и D4

6.4 Агрегированные модели

Затем из отобранных отдельных моделей (табл. 4) составлялись агрегированные модели. Из-за небольшого размера датасетов для агрегации использовалась линейная регрессия с неотрицательными весами.

$$S(d, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \alpha_k S_k(d), \;\; \alpha_k \geqslant 0,$$

где вектор $\boldsymbol{\alpha}$ является решением оптимизационной задачи

$$\sum_{(d,\; d') \in D} \mathcal{L}(S(d', \boldsymbol{\alpha}) - S(d, \boldsymbol{\alpha})) + \lambda \operatorname{Reg}(\boldsymbol{\alpha}) \to \min_{\boldsymbol{\alpha}},$$

где $\mathcal{L}(M)$ — невозрастающая функция отступа M, а Reg — регуляризатор. Для обучения агрегированных моделей использовалось 80% датасета, а для валидации — оставшиеся 20%. В экспериментах сравнивались модели как без регуляризатора, так и с L1-, L2-регуляризаторами или elastic net с гиперпараметром смешения равным 0.5. Гиперпараметр λ перебирался по сетке $[10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1]$. Проверялись следующие функции потерь $\mathcal{L}$ для отступа M:

$\mathcal{L}_1(M) = (1 - M).clip(min = 0), \quad \mathcal{L}_2(M) = |1 - M|, \quad \mathcal{L}_3(M) = (1 - M^2),$
$\mathcal{L}_4(M) = \log(1 + e^{-M}), \quad \mathcal{L}_5(M) = \frac{1}{1+e^M}, \quad \mathcal{L}_6(M) = e^{-M}.$

В таблицах 6, 7 показаны лучшие по точности на валидационных частях соответственно датасетов D4, D3 агрегированные модели, составленные из 21-ой отдельной модели (см. табл. 4) для каждой функции потерь.

В $\mathcal{L}_6(M)$ возникла проблема с переполнением, поэтому её результаты в таблицах 6, 7 не представлены. Также заодно были проверены следующие функции, которые в итоге показали совсем плохое качество, поэтому их результаты в статье не представлены:
$\mathcal{L}_7(M) = -|M|$, $\mathcal{L}_8(M) = -M^2$, $\mathcal{L}_9(M) = 1 - M$, $\mathcal{L}_{10}(M) = (-M)^3$.

№	**Ф-ция потерь**	**Reg**	**λ**	**Точн. на D4 [вал.], %**
1	$\mathcal{L}_1$	L2	10^{-4}	**92.78**
2	$\mathcal{L}_2$	L1	10^{-2}	91.24
3	$\mathcal{L}_3$	L1	10^{-3}	92.14
4	$\mathcal{L}_4$	L1	10^{-3}	88.00
5	$\mathcal{L}_5$	L1	1	82.73

Табл. 6. Точность на вал. части D4 ансамблей из 21 модели для каждой ф-ции потерь

№	**Ф-ция потерь**	**Reg**	**λ**	**Точн. на D3 [вал.], %**
1	$\mathcal{L}_1$	L2	10^{-3}	93.61
2	$\mathcal{L}_2$	L1	10^{-2}	93.31
3	$\mathcal{L}_3$	L2	10^{-3}	**94.51**
4	$\mathcal{L}_4$	L1	0.1	91.11
5	$\mathcal{L}_5$	L1	1	90.81

Табл. 7. Точность на вал. части D3 ансамблей из 21 модели для каждой ф-ции потерь

Из экспериментов было видно, что из всех функций потерь устойчивее всего хорошее качество при подборе гиперпараметров показывали $\mathcal{L}_1(M)$ и $\mathcal{L}_2(M)$, то есть они менее чувствительны к подбору гиперпараметров.

Индекс	**Точн. на D3 [вал.], %**	**Точн. на D4 [вал.], %**
FKGL_{ru}	89.71	88.40
FRES_{ru}	**89.81**	**89.90**
CLI_{ru}	89.11	87.90
SMOG_{ru}	87.31	86.71
ARI_{ru}	89.31	88.75
LIX	89.01	87.76
TI	**89.81**	88.60

Табл. 8. Точность индексов удобочитаемости на вал. частях D3 и D4

Результаты индексов удобочитаемости на аналогичных валидационных частях датасетов D3, D4 представлены в таблице 8.

Изучим теперь подробнее точность агрегированной модели, добившейся наилучшего качества на валидации на D4 (1-ая в табл. 6). Посмотрим на все точности ранжирования фрагментов текстов между отдельными парами учебных классов (см. табл. 9).

Точн.	**6**	**7**	**8**	**9**	**10**	**10+**	**11**	**11+**
5	1	1	1	1	1	1	1	1
6	—	0.95	1	1	1	1	1	1
7	—	—	0.975	1	1	1	1	1
8	—	—	—	0.955	0.97	1	1	1
9	—	—	—	—	0.636	0.953	0.935	1
10	—	—	—	—	—	0.705	0.736	0.98
10+	—	—	—	—	—	—	0.591	0.984
11	—	—	—	—	—	—	—	0.98

Табл. 9. Точности по парам классов на вал. части D4 для ансамбля из 21 модели

Из таблицы 9 можно увидеть, что агрегированная модель хорошо ранжирует по сложности фрагменты текстов из учебных классов, отстоящих более чем на один—два года. Также заметно, что чем младше учебные классы у обоих фрагментов текстов в паре, тем легче для модели их отранжировать, что выглядит логично, так как рост сложности текстов учебников средней школы должен быть заметно более резким, чем у учебников старшей школы.

Посмотрим на аналогичную таблицу точностей для индекса удобочитаемости, показавшего себя лучше других — FRES (табл. 10).

Точн.	**6**	**7**	**8**	**9**	**10**	**10+**	**11**	**11+**
5	1	1	1	1	1	1	1	1
6	—	0.8	1	1	1	1	1	1
7	—	—	0.975	1	1	1	1	1
8	—	—	—	0.736	0.993	1	1	1
9	—	—	—	—	0.882	0.915	0.871	0.991
10	—	—	—	—	—	0.524	0.491	0.967
10+	—	—	—	—	—	—	0.341	0.992
11	—	—	—	—	—	—	—	1

Табл. 10. Точности по парам классов на вал. части D4 для FRES

6.5 Абляционные исследования

В данном эксперименте мы уменьшим число моделей с отдельных уровней в агрегированной модели так, чтобы не ухудшить, а может быть и улучшить качество. Для этого посмотрим на вектор $\boldsymbol{\alpha}$, полученный в результате обучения нашей основной агрегированной модели (1-ая в табл. 6). Отсортируем его компоненты по убыванию: первыми будут идти веса, соответствующие отдельным моделям, вносящим наибольший вклад в определение сложности текста (см. рис. 2).

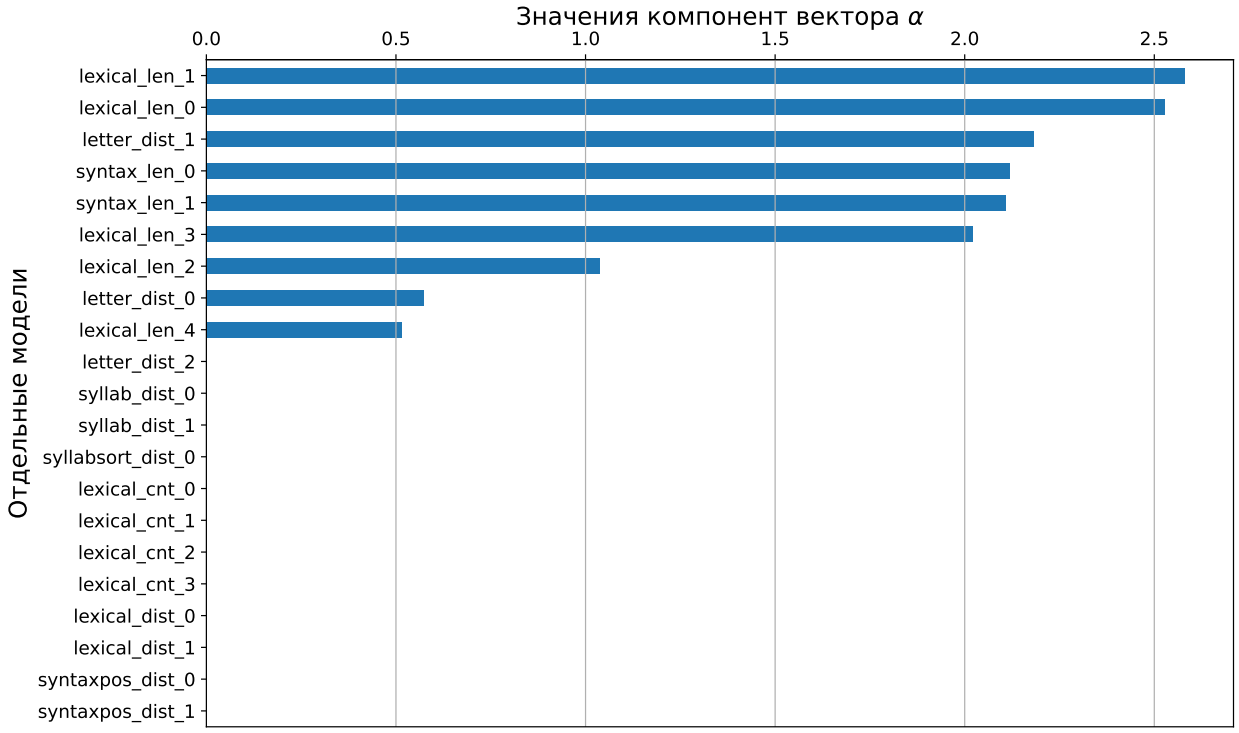

Рис. 2. Значимость отдельных моделей

Далее при проверке качества на валидационной части датасета D4 разных агрегированных моделей с одним удалённым блоком отдельных моделей одного типа выяснилось, что удаление блока лексических сложностных моделей длины или синтаксических сложностных моделей длины приводит к значительной потере качества у всех агрегированных моделей с функцией потерь $\mathcal{L}_1$ и регуляризацией (смотрим на них, поскольку у них самое лучшее качество). Удаление блока фонетических частотных моделей приводит к падению качества на большинстве таких агрегированных моделей. Удаление остальных блоков не приводило к значительным потерям качества, а где-то даже повышало его. На рис. 3 показано изменение точности лучшего по качеству ансамбля при удалении из блока моделей одного типа.

Затем путём тестирования разных наборов блоков в агрегированной модели на валидационной части D4, а также с учётом удаления компонентов блоков с наименьшим вкладом, проводился дальнейший отбор наилучших структур агрегации. В результате

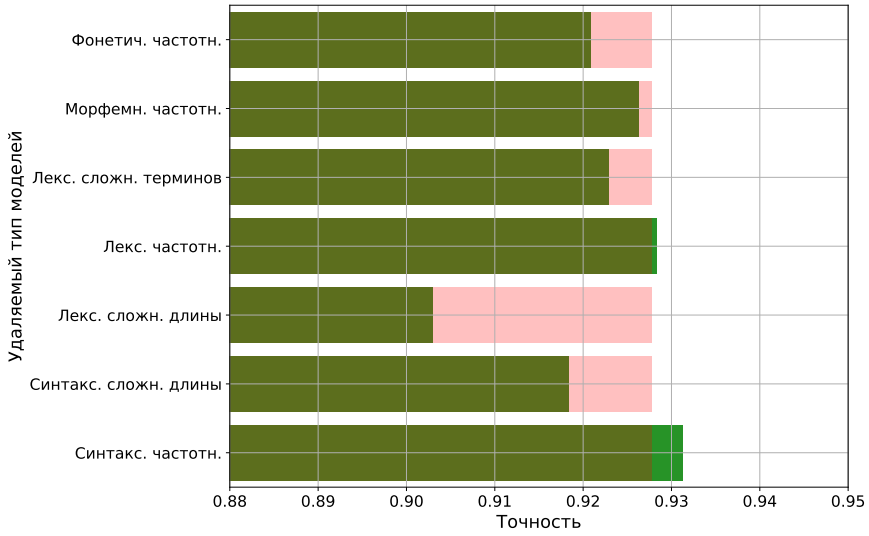

Рис. 3. Изменение точности на вал. части D4 при удалении из ансамбля блока моделей одного типа: розовым показано ухудшение точности относительно ансамбля из 21 модели, ярко-зелёным соответственно — улучшение

была получена следующая агрегированная модель из 9 отдельных моделей, которая показала лучшее качество из опробованных вариантов:

- Фонетические частотные модели: $letter\_dist\_0$, $letter\_dist\_1$;
- Лексические сложностные модели длины: $lexical\_len\_0$, $lexical\_len\_1$, $lexical\_len\_2$, $lexical\_len\_3$, $lexical\_len\_4$;
- Синтаксические сложностные модели длины: $syntax\_len\_0$, $syntax\_len\_1$.

Посмотрим на лучшие точности на валидационных частях датасетов D4, D3 агрегированных моделей для каждой из функций потерь (см. соответственно табл. 11, 12):

№	**Ф-ция потерь**	**Reg**	**λ**	**Точн. на D4 [вал.], %**
1	$\mathcal{L}_1$	elastic net	10^{-4}	**93.48**
2	$\mathcal{L}_2$	L2	10^{-2}	92.33
3	$\mathcal{L}_3$	L2	0.1	92.33
4	$\mathcal{L}_4$	—	0	93.23
5	$\mathcal{L}_5$	—	0	93.33
6	$\mathcal{L}_6$	L1	10^{-3}	93.23

Табл. 11. Точность на вал. части D4 ансамблей из 9 моделей для каждой ф-ции потерь

№	Ф-ция потерь	Reg	λ	Точн. на D3 [вал.], %
1	$\mathcal{L}_1$	—	0	94.91
2	$\mathcal{L}_2$	elastic net	10^{-2}	94.51
3	$\mathcal{L}_3$	—	0	**95.60**
4	$\mathcal{L}_4$	—	0	93.51
5	$\mathcal{L}_5$	L1	10^{-4}	94.61
6	$\mathcal{L}_6$	L2	10^{-4}	94.91

Табл. 12. Точность на вал. части D3 ансамблей из 9 моделей для каждой ф-ции потерь

Из экспериментов было видно, что из всех функций потерь устойчивее всего хорошее качество при подборе гиперпараметров показывали $\mathcal{L}_1(M)$, $\mathcal{L}_2(M)$, $\mathcal{L}_3(M)$ и $\mathcal{L}_6(M)$. А для $\mathcal{L}_4(M)$, $\mathcal{L}_5(M)$ здесь лучше не использовать регуляризацию вовсе, так как с ней качество быстро падает. Также можно заметить, что у нас составлен хороший набор отдельных моделей для агрегации, поэтому почти с любой функцией потерь можно получить приемлемое качество.

Таким образом, из экспериментов получается, что лучше всего использовать функцию потерь $\mathcal{L}_1(M)$ со слабой регуляризацией (любой с гиперпараметром $\lambda = 10^{-4} \dots 10^{-3}$) или без неё.

Посмотрим теперь для наилучшей модели (1-ая в табл. 11) на все точности ранжирования фрагментов текстов между отдельными парами учебных классов (см. табл. 13).

Точн.	**6**	**7**	**8**	**9**	**10**	**10+**	**11**	**11+**
5	1	1	1	1	1	1	1	1
6	—	0.8	1	1	1	1	1	1
7	—	—	0.95	1	1	1	1	1
8	—	—	—	0.918	0.978	1	1	1
9	—	—	—	—	0.773	0.991	1	1
10	—	—	—	—	—	0.699	0.755	1
10+	—	—	—	—	—	—	0.545	0.976
11	—	—	—	—	—	—	—	0.98

Табл. 13. Точности по парам классов на вал. части D4 для ансамбля из 9 моделей

Если сравнить таблицу 13 с таблицей 9, то можно сделать вывод, что ансамбль из 9 моделей лучше справляется с ранжированием фрагментов текста более старших учебных классов. В наших датасетах в старших классах содержится больше фрагментов текста, поскольку учебники старших классов более объёмные. Поэтому повышение качества ранжирования в области старших учебников сильнее оказывает влияние на

общую точность модели. Также следует ещё раз отметить, что сравнивать между собой по сложности тексты из учебников старших классов сложнее, чем из учебников младших.

6.6 Зависимость точности от средней длины фрагмента

В данном эксперименте проведён анализ зависимости точности ранжирования от средней длины фрагмента текста в датасете. Для этого используются все построенные датасеты на основе учебников (см. табл. 3).

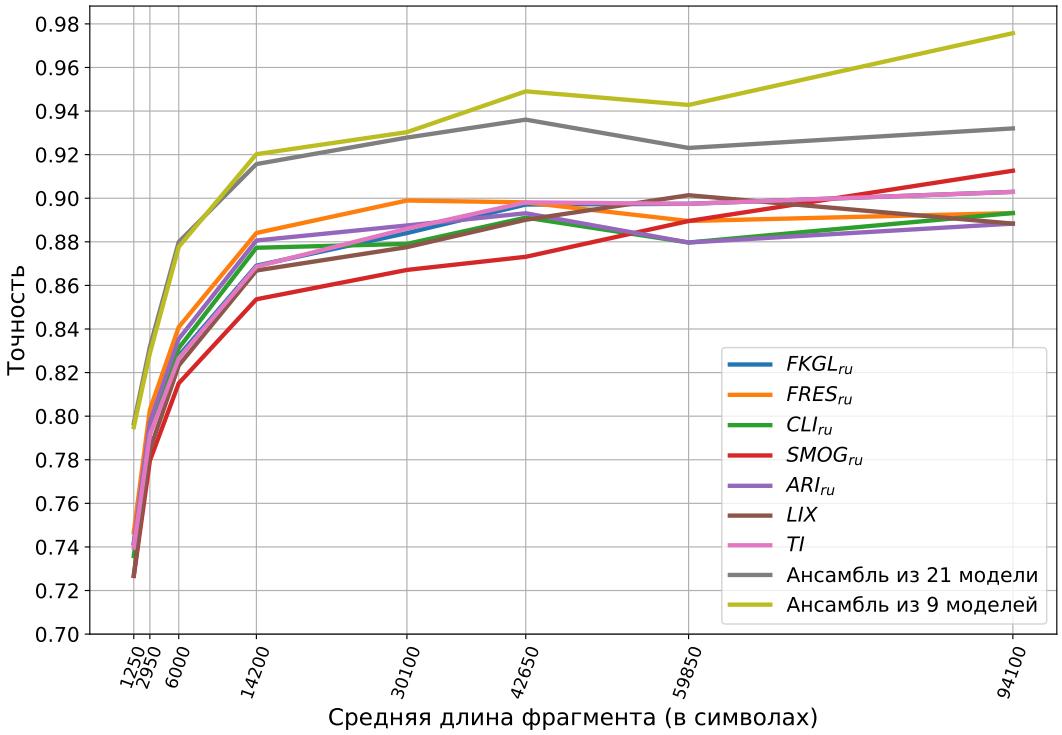

Рис. 4

На рис. 4 можно заметить, что точность при уменьшении длины фрагмента текста начинает снижаться, как у моделей, так и у индексов удобочитаемости. Особенно резкое падение заметно, начиная с длины 14200 символов и меньше. В то время как при длинах бóльших 14200 символов у многих индексов и моделей точность ранжирования выходит плато. Также из рисунка видно, что агрегированные модели демонстрируют качество выше, чем у индексов удобочитаемости, при всех длинах фрагментов текстов, а ансамбль из 9 моделей — выше, чем у ансамбля из 21 модели. Для данного эксперимента был выбран ансамбль из 21 модели с функцией потерь $\mathcal{L}_1(M)$ и с L2 регуляризацией с гиперпараметром $\lambda = 10^{-4}$ и ансамбль из 9 моделей с функцией потерь $\mathcal{L}_1(M)$ без регуляризации.

7 Заключение

В данной работе исследован метод оценивания когнитивной сложности текста, основанный на моделях, использующих квантили. В частности, реализованы модели на фонетическом, морфемном, лексическом и синтаксическом уровнях языка, а также их агрегации. Отдельные модели сложности обучаются по референтному корпусу текстов из Википедии, считая эмпирические распределения оценок сложности для каждого типа токена. Агрегированные модели сложности обучаются по датасетам, сформированным из учебников по обществознанию разных учебных классов. Все рассмотренные модели сравнивались по качеству с индексами удобочитаемости, адаптированными для русского языка. Из отдельных моделей наилучшее качество показали лексические сложностные модели длины слова, превзойдя все индексы удобочитаемости. Агрегация из 21-ой отдельной модели всех типов ещё значительнее превзошла по точности ранжирования пар фрагментов текстов все индексы удобочитаемости. Была изучена её точность для каждой пары учебных классов в отдельности, что позволило убедиться в адекватности построенных экспериментов. Было замечено, что агрегированная модель хорошо ранжирует по сложности фрагменты текстов из учебных классов, отстоящих более чем на один—два года, и что чем младше учебные классы у обоих фрагментов текстов в паре, тем легче для модели их отранжировать правильно. Далее был проведён анализ вклада отдельных компонентов в агрегированную модель (абляционные исследования), в результате которого был найден ансамбль из девяти отдельных моделей, показавший дальнейшее улучшение качества. В него вошли модели следующих типов: фонетическая частотная модель, лексическая сложностная модель длины слова и синтаксическая сложностная модель длины синтаксической связи в предложении. Также в работе был проведён анализ зависимости точности ранжирования от средней длины фрагмента текста в датасете, в результате которого выяснилось, что с уменьшением средней длины фрагмента точность снижается. Особенно резкое падение начинается при количестве символов меньшем 14200.

8 Положения, выносимые на защиту

1) Реализованы квантильные модели оценивания сложности текста на фонетическом, морфемном, лексическом и синтаксическом уровнях языка.

2) Реализованы ансамбли из отдельных моделей и проведён отбор наилучших (абляционные исследования).

3) Подготовлены выборки пар фрагментов текстов на основе учебников по обществознанию разных учебных классов. Исследована точность ранжирования

для каждой пары учебных классов. Изучена зависимость точности от средней длины фрагмента текстов в датасете.

4) Проведено экспериментальное сравнение предлагаемого квантильного подхода с адаптированными для русского языка индексами удобочитаемости. Предлагаемый агрегированный квантильный подход демонстрирует более высокую точность ранжирования пар фрагментов текстов.

Тезисы доклада по данной работе приняты к публикации в сборнике тезисов конференции «Ломоносов-2023» [15]. На момент подготовки текста диссертации статья по ней принята к представлению на конференции «Диалог 2023» с доработками.

Список литературы

[1] Eremeev M. A., Vorontsov K.V. Lexical quantile-based text complexity measure. Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2019, p. 270–275

[2] Senter R. J., Smith. E. A. Automated readability index. AMRL-TR-66-220, 1967.

[3] Björnsson C.-H. Läsbarhet: Lesbarkeit durch Lix. Stockholm: Liber, 1968.

[4] McLaughlin G. H. SMOG Grading — a New Readability Formula. Journal of reading, Vol. 12, No. 8, 1969, p. 639–646

[5] Coleman M., Liau. T. L. A computer readability formula designed for machine scoring. Journal of Applied Psychology, Vol. 60, No. 2, 1975, p. 283–284.

[6] Rudolf Flesch. A new readability yardstick. Journal of Applied Psychology, Vol. 32, No. 3, 1948, p. 221–233.

[7] Kincaid J.P., Fishburne R.P., Rogers R.L., Chissom B.S. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Memphis, TN: Naval Air Station, 1975, p. 40

[8] Тулдава Ю.А. Об измерении трудности текстов. Учен. зап. Тарт. ун-та: Труды по методике преподавания иностранных языков, Вып. 345, 1975, с. 102–120.

[9] Оборнева И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис. ... канд. пед. наук: 13.00.02. — М., 2006. — 165 с.

[10] Бегтин И.В. Plain Russian Language [Электронный ресурс]. — Электрон. дан. — 2014. — URL: https://github.com/infoculture/plainrussian. (дата обращения: 29.04.2023).

[11] Шкарин С. С. ruTS, a library for statistics extraction from texts in Russian [Электронный ресурс]. — Электрон. дан. — 2021. — URL: https://github.com/SergeyShk/ruTS. (дата обращения: 29.04.2023)

[12] Дмитриева А. В. «Искусство юридического письма»: количественный анализ решений Конституционного Суда Российской Федерации // Сравнительное конституционное обозрение. — Вып. 3(118). — изд-во Центр конституционных исследований, 2017. — с. 125–133.

[13] Straka M., Straková J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies, 2017, p. 88–99.

[14] Solovyev V., Ivanov V., Solnyshkina M. Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics. Journal of intelligent & fuzzy systems, Vol. 34, No. 5, 2017, p. 3049–3058

[15] Веселов А.С. Агрегирование квантильных моделей для оценивания когнитивной сложности текста. XXX Международная конференция студентов, аспирантов

и молодых учёных «Ломоносов», секция «Вычислительная математика и кибернетика», 2023, с. 14–15.