ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 60 за 2024 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

М.Ю. Кислицына, Ю.Н. Орлов

Структура ошибок распознавания автора текста методом триграмм

Рекомендуемая форма библиографической ссылки: Кислицына М.Ю., Орлов Ю.Н. Структура ошибок распознавания автора текста методом триграмм // Препринты ИПМ им. М.В.Келдыша. 2024. № 60. 24 с. https://doi.org/10.20948/prepr-2024-60
https://library.keldysh.ru/preprint.asp?id=2024-60

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

М.Ю. Кислицына, Ю.Н. Орлов

Структура ошибок распознавания автора текста методом триграмм

Москва – 2024

Кислицына М.Ю., Орлов Ю.Н.

Структура ошибок распознавания автора текста методом триграмм

Собрана статистика эталонных триграмм для полного корпуса литературных текстов на русском языке, включая переводных зарубежных авторов. На данном корпусе исследована ошибка идентификации автора текста методом ближайшего эталона. С помощью статистики расстояний от автора до эталона корпуса и ширины носителя распределений триграмм были определены классы авторов, для которых ошибка распознавания наибольшая. Кроме того, был получен подкорпус с нулевой ошибкой классификации. На данном подкорпусе рассмотрена ошибка идентификации ошибочно классифицированных текстов, не включенных в правильный подкорпус.

Ключевые слова: триграммы, метод ближайшего соседа, распознавание автора текста, классификация ошибок

Kislitsyna M.Yu., Orlov Yu.N.

The structure of errors in recognizing the author of the text by the trigram method

In this work statistics of trigrams text features for the complete corpus of literary texts in Russian, including translated foreign authors, have been collected. The error of identifying the author of a text using the nearest pattern method has been investigated on this corpus. The classes of authors with large error recognition were determined by statistics of the distances between an author and the corpus feature and the number of non-zero entries in trigram vector. Moreover, a case of creating a subcorpus with zero classification error was obtained. The identification error of incorrectly classified texts that don’t belong to the subcorpus is considered.

Keywords: trigrams, nearest neighbor method, text author recognition, errors classification

Содержание

Введение

Работа посвящена продолжению анализа результатов тестирования метода распознавания автора текста на максимально большом корпусе литературных текстов на русском языке. В этот корпус вошли все доступные в настоящее время литературные произведения на русском языке. Однако, исходя из статистических требований к обучению автоматического распознавания автора текста, таких произведений у каждого автора должно быть достаточно много, поэтому литераторы с малым числом текстов были исключены из рассмотрения. Вычислительный эксперимент и состав корпуса были описаны ранее в [1]. Целью настоящего анализа является определение характерных типов ошибок машинного распознавания автора текста с тем, чтобы понять, какая часть этих ошибок может быть исправлена тем или иным методом коррекции.

Интерес к данной теме обусловлен распространением технологии искусственного интеллекта на различные области человеческой жизнедеятельности. Машинная обработка текстов на естественных языках представляет одно из важных направлений применения этой технологии для задач классификации текстов и массивов документов [2-4]. Существует большое количество компьютерных программ для их обработки. Все они используют методы статистики для автоматического определения языка, жанра или тематики, а также автора некоторого достаточно большого текста.

При этом применяются два основных подхода (см. [5, 6]). Первый состоит в экспертном отборе определенных текстов для обучения манере написания фраз, т.е. основан на мнении эксперта относительно правил написания произведения определенного формата. Ограничение этого подхода по точности связано с тем, что, во-первых, существуют разные экспертные мнения, и, во-вторых, автор может менять манеру письма.

Второй подход чисто статистический. В нем есть определенная претензия на объективность в том смысле, что методом может воспользоваться любой исследователь, и, независимо от личной точки зрения, всегда будет получать один и тот же результат применительно к одному и тому же набору документов. В основном статистические методы основаны на анализе наиболее часто употребляемых слов, знаков препинания, букв, абзацев и иных чисто технических характеристик текста. Неполная объективность метода для задачи идентификации автора состоит в том, что, вообще говоря, автору нельзя запретить использовать те слова лексикона, которые использует и другой автор. В этом смысле статистика букв более объективна, чем статистика слов. Удивительно, что эта статистика, не отражающая смысловой процесс построения фраз, тем не менее в случае с профессиональными писателями обладает авторским своеобразием.

В обзорной работе [7] сравнивается эффективность чисто статистических методов анализа, основанных на подсчете формальных показателей, таких как

число букв, слов, знаков препинания и т.п., с экспертными методами анализа авторского стиля, оборотов речи, использования литературных приемов. Авторы [7] приходят к выводу, что хотя для литературоведов более ценен экспертный метод, он не обладает достаточной точностью на большом корпусе текстов и, что более существенно, не может быть адекватно реализован в виде формальной компьютерной программы. В то же время статистика букв или буквосочетаний, хотя и не имеет непосредственного литературного смысла, может быть вполне однозначно сопоставлена каждому тексту с указанием погрешности в рамках формальных критериев. Тем самым в контексте задачи машинного распознавания атрибутов текстов статистический метод более эффективен, т.е. имеет меньшую ошибку, чем экспертный.

1. Алгоритм коррекции ошибки распознавания автора текста

Ранее в работах [8, 9] был сформулирован алгоритм коррекции ошибки распознавания автора текста внутри заданной библиотеки эталонов. Одним из видов ошибки является нахождение ближайшего эталона для текста, правильный авторский эталон которого отсутствует. При кросс-валидации в эту категорию попадают тексты, написанные автором в несколько ином стиле, чем большинство его произведений. Этим объясняется корректность вердикта «автор отсутствует в библиотеке», поскольку текст написан как бы иной ипостасью автора. Подчеркнем, что пока мы не подразделяем автора на его отдельные «лица», поскольку автоматически сделать это проблематично, а экспертное мнение в этом случае не объективно.

Предлагаемый алгоритм коррекции ошибки основан на гипотезе устойчивости «своего» эталона и неустойчивости «чужого» при разделении текста на фрагменты. Этот подход является логическим продолжением гипотезы об идентификации автора текста через близость к его эталонному распределению буквосочетаний – в данном случае триграмм. Предполагается, что автор литературного текста, то есть профессиональный писатель, обладает определенным стилем подбора слов для выражения своих мыслей, так что этому стилю отвечает некоторая генеральная совокупность вероятностей буквосочетаний. Отдельный текст в такой трактовке является конечной выборкой из генеральной совокупности.

Если имеется набор таких распределений для разных авторов, то с формальной стороны задача идентификации автора сводится к оценке вероятности правильной идентификации выборок из разных генеральных совокупностей в зависимости от близости между этими совокупностями и длины выборки.

По результатам обработки текстов мы строим эмпирическое распределение $D_a^i(j)$ частот символа j в i-м тексте автора a. Обозначим через N_a^i число символов в данном i-м тексте. Пусть также n_a – число произведений автора a. Тогда эмпирической оценкой эталона $F_a(j)$ автора a является

взвешенное распределение частот по совокупности всех текстов, достоверно принадлежащих данному автору:

$$F_a(j) = \frac{1}{N_a}\sum_{i=1}^{n_a} N_a^i D_a^i(j), \quad N_a = \sum_{i=1}^{n_a} N_a^i . \tag{1}$$

Расстояние между текстами, а также между текстом и эталоном понимается в смысле расстояний между соответствующими распределениями в норме L1. Тогда расстояние от текста до эталона корпуса дается формулой

$$z_{ab}^i = \frac{1}{1-\delta_{ab} N_b^i / N_b}\sum_{j=1}^{J}\left|D_a^i(j) - F_b(j)\right|, \tag{2}$$

где δ_{ab} – символ Кронекера. Формально для безошибочного распознавания автора текста требуется, чтобы

$$\forall i,a,b:\ z_{aa}^i < z_{ab}^i,\ b \neq a . \tag{3}$$

Ошибкой идентификации будем называть долю текстов, автор которых был неверно определен по формуле

$$a = \arg\min_b z_{ab} . \tag{4}$$

Формально ошибка определяется как отношение числа нарушений неравенства (3) к общему числу текстов.

Пусть тестируется достаточно длинный текст, который может быть разрезан на некоторое количество фрагментов. Уменьшая длину фрагмента текста, мы увеличиваем тем самым статистическую неопределенность оценивания выборочных частот, но увеличиваем число экспериментов по идентификации автора одного и того же текста. В работе [8] было сформулировано подтверждающее правило идентификации: если при разбиении текста на 2^n фрагментов равных длин оказывается, что для каждого значения $4 \geq n \geq 1$ хотя бы один фрагмент имеет авторство, отвечающее нулевому разбиению, т.е. полному тексту, то автор найден правильно. Если же хотя бы для одного n таких фрагментов не нашлось, считается, что автор исходного текста был опознан неверно. Тестирование этого правила на относительно небольшом корпусе в 100 авторов и примерно 2 тыс. текстов показало, что ошибка может быть уменьшена примерно на 50%. Представляет практический интерес проверить этот подход на полном корпусе русскоязычных литературных текстов.

Эффективность этого метода коррекции определяется сопоставлением между собой ошибок первого и второго родов. Пусть исходный корпус текстов распознан с ошибкой ε. Тогда метод коррекции считается улучшающим, если ошибка α ложного отрицания правильно распознанного текста ухудшает итоговое значение ошибки на величину, меньшую, чем улучшение ошибки за счет вероятности β правильного отрицания неверно распознанного текста. Это означает, что должно быть $\alpha(1-\varepsilon) < \beta\varepsilon$, т.е. $\varepsilon > \dfrac{\alpha}{\alpha+\beta}$.

В нашем случае $\varepsilon = 0,24$, $\alpha = 0,15$, $\beta = 0,85$, так что указанное соотношение выполнено. Анализ показал, что данный метод приводит к снижению ошибки по полному корпусу до величины $\varepsilon' = 0,11$. Сам по себе этот результат представляет большой интерес, поскольку удалось построить независимый индикатор, вероятность срабатывания которого на разных подкорпусах различна. В то же время надо исследовать, с чем связаны успехи и неудачи данного алгоритма. С этой целью мы рассмотрим, какие конкретно подгруппы авторов распознаются лучше или хуже и на каких подгруппах подтверждающий индикатор срабатывает наиболее эффективно.

Одним из параметров, по которому можно автоматически кластеризовать авторов, является расстояние от авторского до среднего эталона полного корпуса, который мы будем далее называть центром. Центр – это литературная проекция лексикона, который не имеет индивидуального своеобразия, а отражает общую статистику языка. Интересно провести анализ того, какие авторы находятся ближе к центру, а какие – дальше, и как это положение сказывается на точности идентификации.

Также имеет смысл рассмотреть зависимость ошибки идентификации от величины носителя эмпирического распределения автора. С одной стороны, чем ближе эталон автора к центру корпуса, тем шире у него носитель. Обратное же, вообще говоря, неверно. Поэтому ширина носителя – число ненулевых триграмм в эталоне – может служить дополнительным параметром, по которому следует классифицировать ошибку.

Заметим, что эффективность применения индикатора может зависеть от того, насколько много текстов у автора, а также и от того, какая доля этих текстов оказалась ошибочно распознанной методом ближайшего соседа. Очевидно, что применение индикатора к авторам, распознанным безошибочно, может только ухудшить идентификацию. И напротив, для авторов, все тексты которых распознаны неверно, применение индикатора будет приводить к улучшению.

2. Зависимость ошибки от расстояния до центра корпуса

Итак, исследовался корпус текстов на русском языке, состоящий из 108 518 текстов, которые написаны 8 287 писателями. Из них 5 084 автора (61% авторов) – русскоязычные, остальные 3 203 автора (39%) рассматриваются в переводах. По количеству литературных произведений состав корпуса следующий: имеется 66 215 текстов на изначально русском языке, что также составляет 61% от общего числа текстов корпуса, а 42 303 текста – переводные. Полный список авторов и текстов корпуса представлен в [10]. Далее именно этот набор произведений будет называться «Полный Корпус» или просто корпус.

Общий объем анализируемых произведений составляет примерно 34 млрд знаков, из них 20 млрд знаков соответствуют текстам на изначально русском языке. Таким образом, русскоязычные и переводные произведения как по

суммарной длине, так и по количеству текстов и количеству авторов соотносятся примерно как 60:40.

Точность идентификации автора методом ближайшего эталона триграмм оказалась равной $\varepsilon = 0,238$. При этом русскоязычные авторы были распознаны с ошибкой $\varepsilon_1 = 0,196$, а переводы – с ошибкой $\varepsilon_2 = 0,301$. Предположительно, такое соотношение показывает, что переводные тексты идентифицируются хуже, чем написанные на родном языке.

Для общего представления о статистике текстов и эталонов приведем ниже на рис. 1-3 характерные распределения расстояний в норме L1.

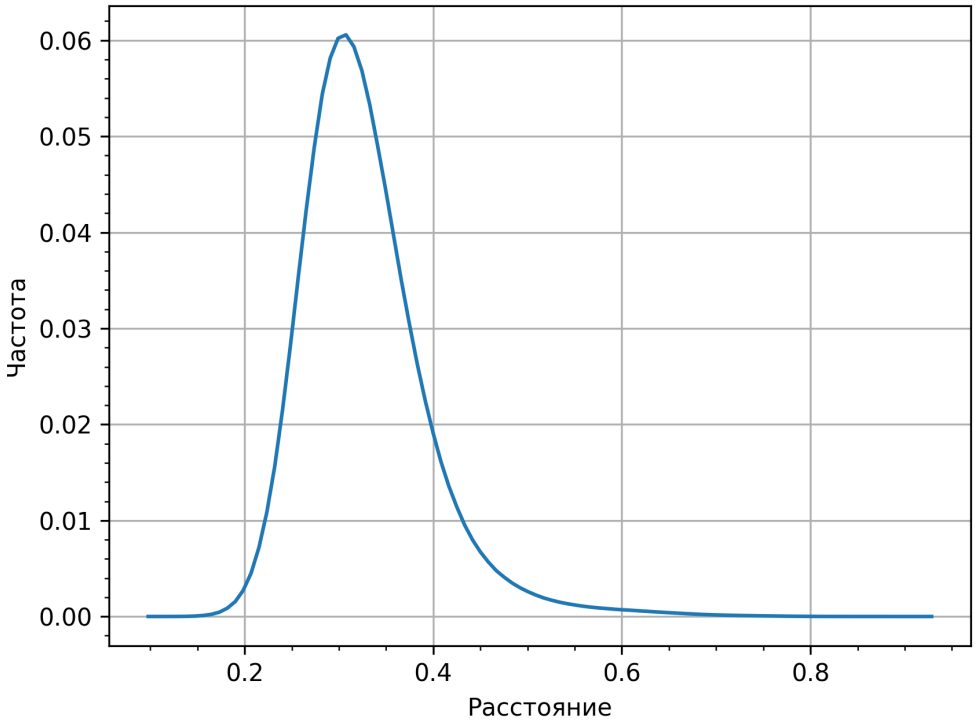

Рис. 1. Распределение расстояний в L1 между эталонами

Таким образом, характерное расстояние между эталонами триграмм составляет величину 0,3. Однако есть и достаточно малые расстояния, порядка 0,15. Это означает, что при большом количестве авторов может возникнуть ситуация, когда расстояние от текста до своего эталона случайно (а возможно, и не случайно) окажется больше, чем до чужого. Это приведет к ошибке идентификации. Задачей анализа ошибок является выяснение того, в каких группах авторов, которые могут быть выделены автоматически, ошибки больше или меньше средней по корпусу.

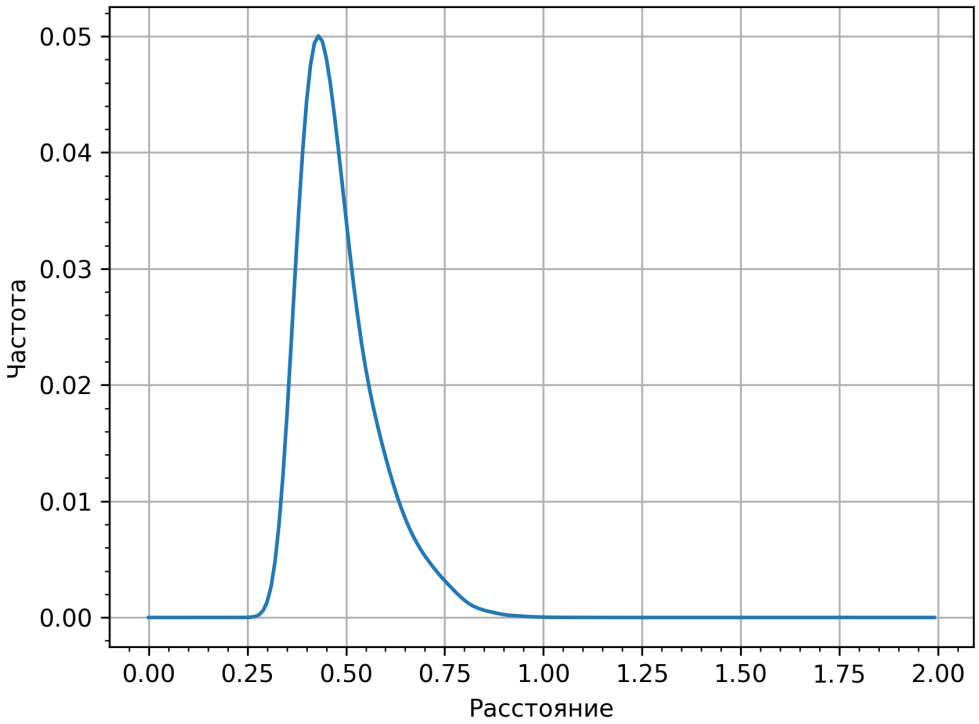

Рис. 2. Распределение расстояний в L1 между текстами

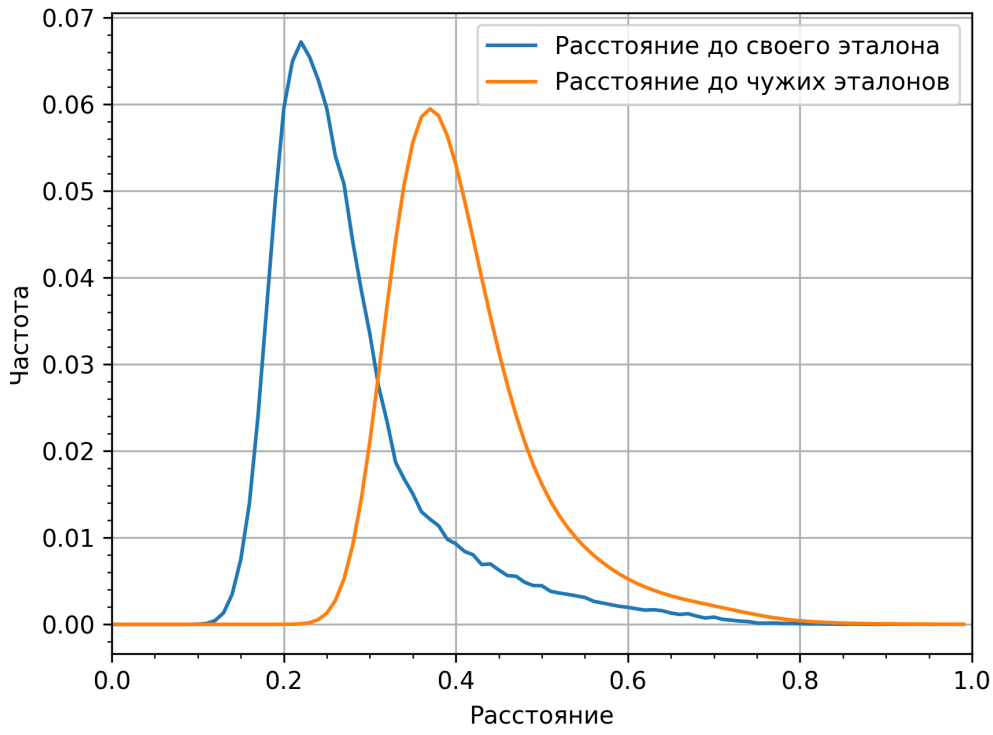

Рис. 3. Распределение расстояний в L1 между текстами и эталонами

Из этих графиков следует, что расстояния между эталонами находятся в промежутке от 0,15 до 0,65, тогда как расстояния между текстами – в промежутке от 0,30 до 0,90. Это означает, что эталоны более информативны,

чем отдельные тексты, и именно по близости к ним, а не к отдельным текстам, следует идентифицировать автора. Это положение находится в согласии с гипотезой о том, что автор пишет текст как выборку из своей генеральной совокупности.

График на рис. 3 показывает, что расстояние от текста до чужого эталона значимо больше, чем до своего. Следовательно, метод распознавания автора по близости к эталону триграмм является статистически адекватным. В то же время надо признать, что литературного обоснования этому мы не имеем. Буквосочетания не являются смысловым признаком, они лишь следуют за ним. То, что по ним можно опознать автора, является чисто эмпирическим фактом, не связанным напрямую с процессом литературного творчества.

На рис. 4 приведен график, показывающий, как ошибка распознавания автора зависит от места авторского эталона по отношению к среднему распределению триграмм корпуса текстов библиотеки из 8287 авторов. По оси абсцисс указано расстояние от эталона автора до центра корпуса. Для каждого классового интервала выявляются авторы, которые в него попадают, после чего подсчитывается суммарное число неверно идентифицированных текстов этих авторов. Соответствующая доля ошибок по отношению к числу текстов, попавших в данный интервал, указывается по оси Y. Для сравнения приведено также распределение расстояний от авторских эталонов до корпуса.

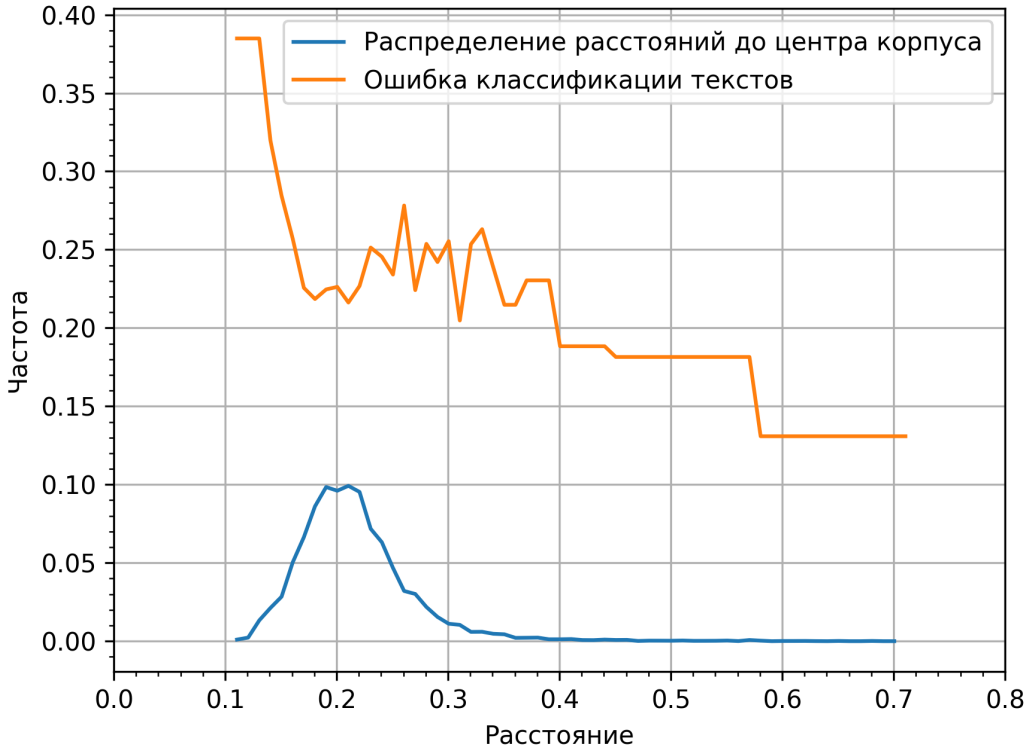

Рис. 4. Распределение ошибки в зависимости от расстояния эталона до центра корпуса

Из сравнения графиков на рис. 4 можно сделать следующие выводы.

Во-первых, авторы, чьи эталоны находятся от центра корпуса на расстоянии, большем 0,4, распознаются методом триграмм с ошибкой порядка 13%. Это в три раза лучше, чем распознавание авторов, находящихся в первом персентиле близости к центру корпуса. Следовательно, авторы с выраженной индивидуальностью творчества распознаются в среднем лучше, чем авторы, похожие на большинство. Однако величина ошибки, равная 13%, указывает на неоднозначность природы ошибки.

Во-вторых, основное число авторов находится на расстоянии от 0,17 до 0,23 до центра корпуса, именно для них распознавание имеет устойчивый локальный минимум порядка 22%. Далее вплоть до расстояний 0,4 распознавание носит колебательный характер от 20 до 27% в среднем на уровне 24%.

Какие авторы находятся на большом расстоянии от центра корпуса? Примерно половину последнего дециля составляют авторы самиздата, чьи тексты не сохраняют индивидуальность распределений триграмм. Это отчасти согласуется с мнением лингвистов о том, что буквосочетания не являются параметром, позволяющим кластеризовать блогосферу, поскольку они выражают свойства лексикона в целом. Тем не менее, профессиональные писатели отличаются от «обычных» тем, что, видимо, работают над собственным стилем, что и приводит к индивидуализации распределений. Что касается другой половины, то это в основном философы с мировым именем: Аристотель, Геродот, Платон, Плутарх, Ницше, Хайдеггер, Кант, Юнг, Фрейд, Рассел, Кьеркегор, Шопенгауэр, Франк, Гегель. Ошибка распознавания текстов этой группы варьируется от нулевой до 100%. Абсолютно без ошибок в этой группе распознано 432 текста 50 авторов из 132 или примерно 38% от авторов данной группы. Всего по корпусу, состоящему из 8287 авторов, безошибочно распознано 2191 авторов, то есть около 26%. Поэтому далекие от центра корпуса авторы распознаются с нулевой ошибкой чаще, чем остальные.

Близкие к центру корпуса авторы составляют совершенно другую картину. Это в основном популярные сериальные писатели с числом произведений в несколько десятков, а иногда и сотен. Без ошибок в первом персентиле расстояний распознано только 8 авторов из 484 (против 50 в последнем персентиле). В целом, как видно из рис. 4, ошибка распознавания в этой области составляет примерно 38%.

Полностью не распознанных авторов в полном корпусе имеется 472. В общем числе их не особенно много, но метод распознавания по распределению буквосочетаний оказался к ним полностью не применимым. Сами эти авторы не имеют какой-то специфической литературной кластеризации. Среди них есть как мастера прозы, так и условно «не особенно известные» авторы. Отметим, что плохо распознаются поэты, которые иногда пишут прозу: это, например, А. Пушкин, И. Гете, А. Блок, И. Бродский, Б. Окуджава, Ю. Визбор. Единственное, что характерно для этой группы авторов – близость к центру корпуса, а не к его периферии.

3. Зависимость ошибки от ширины носителя распределения триграмм

Исследуем теперь зависимость ошибки распознавания автора от величины носителя распределения триграмм.

Рассмотрим сначала ширину носителя авторских эталонов. Из почти 33 тысяч возможных буквосочетаний триграмм имеется довольно большое количество нулевых значений в силу особенностей словообразования в русском языке. Тем не менее, поскольку пробелы исключены, могут появляться довольно неожиданные сочетания типа «ттт» и т.п. На рис. 5 приведена зависимость ошибки распознавания текстов автора от ширины носителя авторского эталона. Этот график был построен следующим образом. Область изменения носителя триграмм от примерно 7 тыс. до 21 тыс. была разделена на классы, в которые объединялись авторы по величине носителя эталона. Далее подсчитывалась ошибка идентификации текстов по каждой группе авторов.

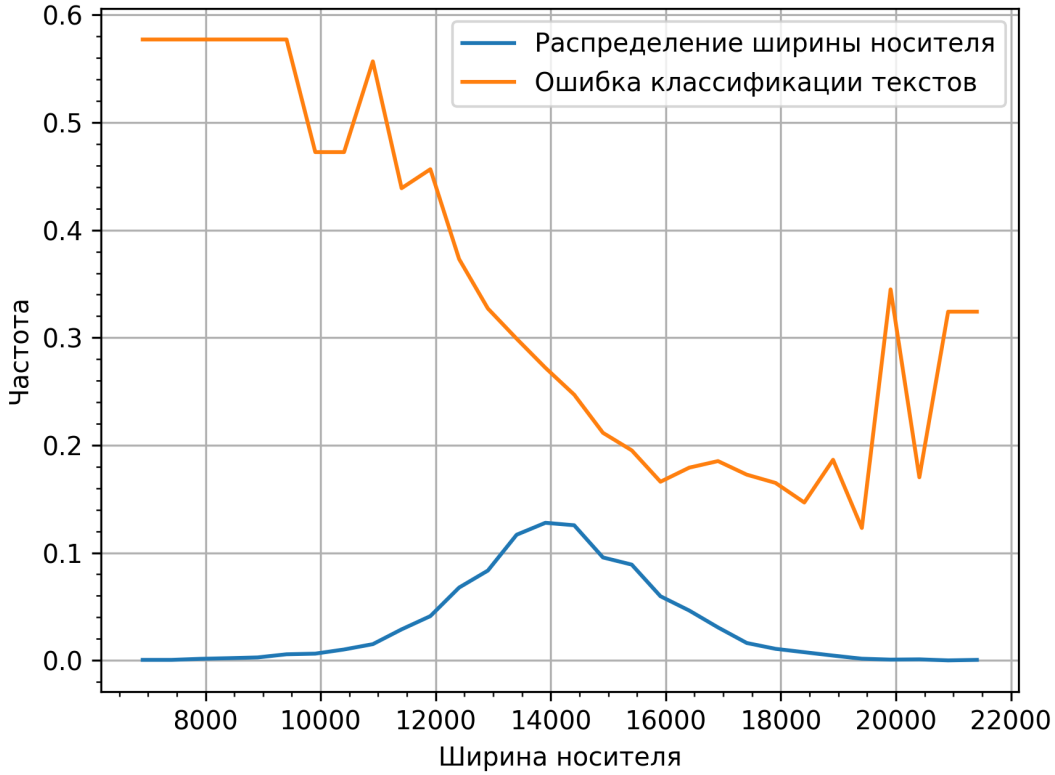

Рис. 5. Распределение ошибки в зависимости от ширины носителя эталона

Из рис. 5 видно, что распределение авторов по ширине носителя эталона приблизительно нормально со средним 14 тыс. и среднеквадратичным отклонением 3 тыс.

Что касается ошибки распознавания текстов соответствующих авторов, то можно выделить три диапазона. Первый характеризуется малостью среднего носителя. Это может быть связано как с малым объемом текста, так и с небольшим разнообразием лексикона. В области от 7 до 11 тыс. ошибка распознавания имеет величину порядка 0,6, т.е. такие авторы пишут, не

сохраняя индивидуального своеобразия. Далее вплоть до примерно 19 тыс. ошибка монотонно снижается с увеличением носителя. Отчасти это можно интерпретировать как возрастание писательского мастерства: чем шире диапазон используемых символов, тем более устойчив авторский стиль. Однако такой простой «критерий мастерства» опровергается последним диапазоном от 19 до 21 тыс. Там ошибка идентификации неожиданно вновь возрастает, имея в то же время немонотонное поведение – наблюдаются колебания от 0,15 до 0,35.

В первом дециле распределения ошибки содержится примерно 500 авторов с носителем от 7 до 11 тыс. В нем находятся в основном русскоязычные «неизвестные» авторы с относительно небольшим (порядка 10) числом произведений. Их доля в первом дециле составляет около 80%. Из «известных» писателей в этой области находятся: Всеволод Гаршин, Михаил Погодин, Павел Бажов, Акутагава Рюноскэ, Адольфо Биой Касарес, Вильгельм Гауф, Пер Лагерквист. Надо отметить, что здесь нет ни одного писателя-классика с «мировой известностью». Что касается зарубежных писателей, то, возможно, имеется влияние перевода. Заметим, что большинство переводных авторов характеризуется очень широким носителем триграмм. Это, в свою очередь, может быть объяснено стараниями квалифицированных переводчиков.

Следует оговориться, что применительно к писательской области деятельности авторы данной работы высказывают лишь свои собственные соображения, касающиеся комментариев или объяснений того или иного эффекта в ошибках распознавания. Мы также не имеем возможности выяснить у конкретного автора, чем обусловлено отклонение какого-либо его текста от эталона, да и такое объяснение будет заведомо носить субъективный характер.

В последнем дециле в диапазоне от 17 до 22 тыс. находятся в основном «сериальные писатели», имеющие несколько десятков произведений. Интересно отметить, что «мировые классики» русской литературы практически отсутствуют и здесь. Однако стараниями переводчиков преобладают переводные писатели, многие из которых с мировым именем. Приведем соответствующий список, упорядоченный по увеличению носителя.

Курт Воннегут, Джон Апдайк, Герберт Уэллс, Эдгар Уоллес, Эмиль Золя, Харуки Мураками, Роберт Шекли, Синклер Льюис, Майн Рид, Энн Райс, Лайон Спрэг де Камп, Уильям Теккерей, Эдгар Берроуз, Артур Конан Дойль, Брайан Ламли, Джоан Роулинг, Эрих Мария Ремарк, Ю Несбё, Роберт Асприн, Джеймс Клавелл, Оноре де Бальзак, Лоис Буджолд, Картер Браун, Дик Фрэнсис, Джуд Деверо, Колин Маккалоу, Чарльз Диккенс, Роберт Силверберг, Анджей Сапковский, Джон Гришем, Рекс Стаут, Гарри Гаррисон, Анри Труайа, Луи Буссенар, Пэлем Вудхауз, Роберт МакКаммон, Жорж Сименон, Айзек Азимов, Джон Карр, Джеймс Чейз, Андрэ Нортон, Глен Кук, Филипп Дик, Эрл Гарднер, Майкл Муркок, Роберт Хайглайн, Алан Фостер, Роберт Ладлэм, Уилбур Смит, Дин Кунц, Терри Пратчетт, Александр Дюма, Дэвид Вебер.

Русскоязычные писатели в последнем дециле следующие: Татьяна Полякова, Вадим Проскурин, Виктор Пелевин, Юрий Поляков, Игорь Пронин, Владимир Сорокин, Виктор Астафьев, Александр Солженицын, Кир Булычев, Михаил Веллер, Виктор Доценко, Эдуард Тополь, Алексей Пехов, Ольга

Володарская, Полина Дашкова, Олег Дивов, Александра Маринина, Андрей Ливадный, Валентин Пикуль, Юрий Власов, Татьяна Устинова, Владимир Васильев, Еремей Парнов, Сергей Лукьяненко, Сергей Алексеев, Ольга Романовская, Далия Трускиновская, Александр Абердин, Александр Мазин, Дарья Калинина, Виктория Платова, Олег Рой, Вера Камша, Елена Хаецкая, Чингиз Абдуллаев, Леонид Влодавец, Кирилл Казанцев, Михаил Серегин, Владимир Поселягин, Андрей Валентинов, Александр Розов, Александр Бушков, Александр Тамоников, Василий Головачев, Фридрих Незнанский, Василий Аксенов, Владислав Крапивин, Борис Акунин, Сергей Зверев.

Из последнего перечня авторов можно сделать вывод о том, что ширина носителя эталона – не вполне информативный параметр. Во-первых, известных писателей-классиков в этом списке почти нет. Во-вторых, большое количество текстов у одного автора может привести к тому, что каждый отдельный текст имеет не очень большой носитель, но в силу разнообразия (если таковое имеется) эталон получается весьма широкий. Именно с этим может быть связана колеблющаяся ошибка последнего дециля. Часть писателей имеет устойчивые триграммы, причем каждый текст достаточно широк в смысле носителя, а часть – меньший носитель и неустойчивые распределения текстов. В этой связи интересно посмотреть на зависимость ошибки распознавания от средней величины носителя текста писателя. Соответствующие графики приведены на рис. 6.

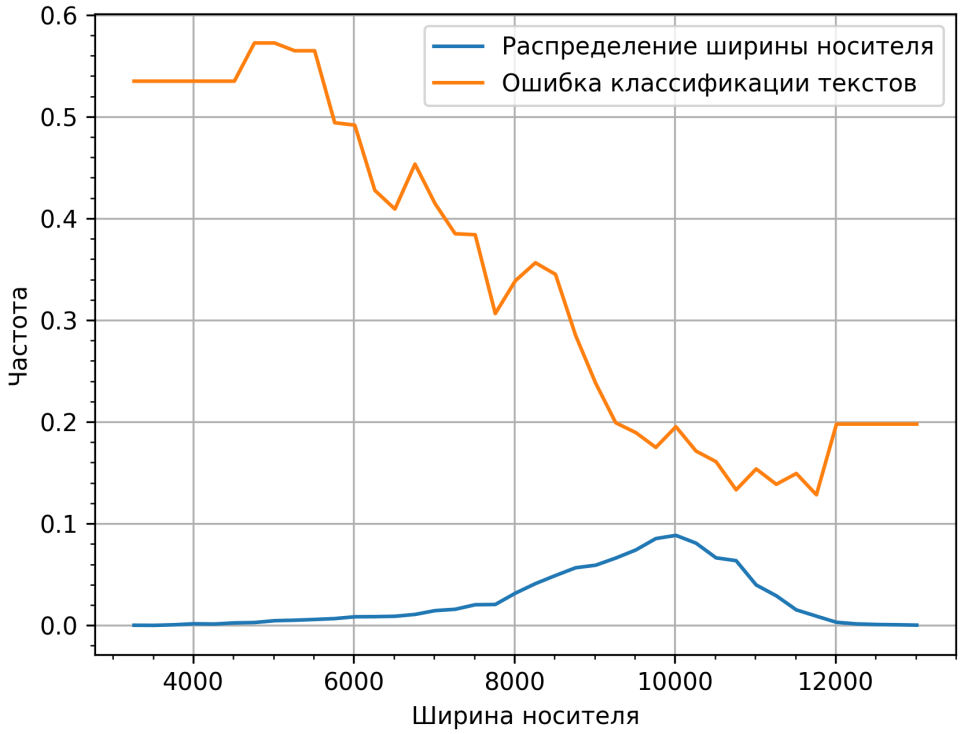

Рис. 6. Распределение ошибки в зависимости от средней ширины носителя

Как видно, здесь распределение авторов по средней ширине носителя текста заметно отличается от гауссовского. Носители отдельных текстов лежат

в диапазоне от 3 тыс. до примерно 13 тыс. Наиболее часто произведения имеют носитель в диапазоне 8-11 тыс.

Как и на рис. 5, здесь мы также имеем три области: малых средних носителей текста от 3 до 8 тыс., промежуточных от 8 до 11 тыс. и больших от 11 до 13 тыс. В целом можно уверенно сказать, что в диапазоне от 5 до 11 тыс. ошибка идентификации монотонно снижается с ростом носителя. Первый и последний децили распределения ошибки представляют основной интерес, поскольку они выпадают из общей приблизительно монотонной картины.

Первый дециль распределения среднего носителя имеет диапазон от 3 до 7 тыс. Он состоит примерно на 80% из русскоязычных авторов, как и первый дециль распределения носителя эталона. Однако теперь в этом дециле гораздо больше известных авторов, многие из которых относятся к мировым классикам. Это означает, что ширина носителя может статистически отображать мастерство автора, но все же не является критерием значимости писателя. Всего лишь 5% русских писателей из списка первого дециля относится к условной категории «известных». По возрастанию среднего носителя текста эти авторы следующие.

Николай Карамзин, Константин Циолковский, Дмитрий Фурманов, Михаил Погодин, Саша Черный, Валентин Пикуль, Григорий Остер, Иван Бунин, Николай Гоголь, Павел Бажов, Александр Бестужев-Марлинский, Анатолий Алексин, Евгений Замятин, Дмитрий Мамин-Сибиряк, Александр Грин, Юрий Олеша, Кир Булычев.

Из переводных авторов в первом дециле находятся: Рэй Бредбери, Ганс Андерсен, Вильгельм Гауф, Эдгар По, Рюноксэ Акутагава, Роберт Блох, О’Генри, Вашингтон Ирвинг, Боэций, Адольфо Биой Касарес, Брет Гарт, Туве Янсон, Платон, Квинт Тертуллиан, Аристотель, Редьярд Киплинг, Ги де Мопассан, Уильям Сароян.

Последний дециль среднего носителя текста от 11 до 13 тыс. содержит список авторов, значительно отличающийся от списка авторов соответствующего дециля носителя эталона. Теперь в нем почти полностью отсутствуют сериальные писатели (за исключением Корецкого), да и вообще как известных, так и просто «популярных» авторов очень мало. Как русскоязычных, так и переводных известных писателей порядка 2%. Перечислим их в порядке увеличения среднего носителя.

Даниил Корецкий, Павел Загребельный, Мария Арбатова, Еремей Парнов, Полина Дашкова, Зинаида Гиппиус, Вилис Лацис, Владимир Орлов, Александр Розов, Юрий Власов.

Зарубежные авторы: Лион Фейхтвангер, Джон О’Хара, Марсель Пруст, Уильям Стайрон, Джон Ле Карре, Гор Видал, Колин Маккалоу, Эрик ван Ластбадер, Дэн Браун, Питер Страуб, Уилбур Смит.

Тем не менее в основной массе текстов увеличение носителя приводит к повышению устойчивости идентификации автора.

4. Безошибочно определяемый подкорпус авторов

Очень важным аспектом модели идентификации является то, что существует большой подкорпус авторов, оставшиеся тексты которых распознаются абсолютно без ошибок. Для его построения из полного корпуса были удалены все тексты, которые имели ошибочную идентификацию автора. Если при этом у автора оказывалось менее пяти текстов, он полностью удалялся из корпуса. За четыре итерации корпус был очищен от ошибок. Итоговый результат представляет собой «Идеальный корпус», состоящий из 76661 текстов, написанных 5871 авторами. Следовательно, предлагаемый метод является правильным в весьма представительной части практического приложения.

При построении идеального корпуса были исключены:

- - 460 авторов (из 1457), имеющих 5 текстов;
- - 369 авторов (из 1058), имеющих 6 текстов;
- - 247 авторов (из 830), имеющих 7 текстов;
- - 134 автора (из 650), имеющих 8 текстов;
- - 137 авторов (из 543), имеющих 9 текстов;
- - 684 автора (из 2473), имеющих 10 текстов;
- - 138 авторов (из 735), имеющих 11-20 текстов;
- - 63 автора (из 256), имеющих 21-30 текстов;
- - 24 автора (из 91), имеющих 31-40 текстов;
- - 46 авторов (из 155), имеющих 41-50 текстов;
- - 10 авторов (из 30), имеющих 51-100 текстов.

Следовательно, в целом сокращение шло в равномерных пропорциях по писателям – в том смысле, что оно было примерно одинаково как для авторов с малым числом произведений, так и с большим. Проведенную фильтрацию можно рассматривать как завершение этапа машинного обучения. В результате построен базис эталонов писателей, который работает с нулевой ошибкой на самом себе. Далее можно решать инженерную задачу – об определении в этом базисе новых текстов тех же самых авторов. В случае если они определяются неверно, этот вариант трактуется как появление «второго лица» автора.

Приведем неполный список «известных» в нашем понимании авторов, которые вошли в правильный корпус. Русскоязычные авторы следующие (указано также количество безошибочно определяемых текстов):

Дарья Донцова: 264, Владимир Колычев: 221, Дарья Калинина: 216, Фридрих Незнанский: 178, Чингиз Абдуллаев: 174, Сергей Зверев: 158, Юрий Никитин: 146, Сергей Самаров: 139, Александр Тамоников: 123, Василий Головачёв: 112, Елена Арсеньева: 111, Александр Афанасьев: 109, Светлана Алешина: 104, Хайдарали Усманов: 104, Владимир Поселягин: 98, Татьяна Полякова: 96, Галина Романова: 95, Владислав Крапивин: 85, Кир Булычев: 81, Андрей Ливадный: 80, Александр Бушков: 73, Елена Хаецкая: 69, Дмитрий Емец: 65, Александра Маринина: 63, Влада Ольховская: 63, Юрий Корчевский: 63, Кира Стрельникова: 62, Алекс Орлов: 62, Михаил Март: 62, Николай Лесков: 51, Татьяна Устинова: 49, Виктория Токарева: 48, Олег Рой: 48,

Владимир Михайлов: 46, Дмитрий Мамин-Сибиряк: 45, Ник Перумов: 45, Далия Трускиновская: 38, Сергей Лукьяненко: 38, Борис Акунин: 37, Кирилл Казанцев: 37, Сергей Снегов: 35, Сергей Алексеев: 34, Николай Леонов: 33, Александр Проханов: 33, Александр Грин: 32, Глеб Успенский: 32, Константин Станюкович: 31, Александр Мазин: 30, Дина Рубина: 30, Анатолий Алексин: 29, Андрей Валентинов: 29, Анна Велес: 29, Александр Беляев: 28, Александр Розов: 27, Виктор Пронин: 27, Святослав Логинов: 27, Юрий Нагибин: 27, Александр Казанцев: 26, Максим Горький: 26, Николай Бердяев: 26, Олег Дивов: 26, Василий Звягинцев: 25, Вера Камша: 25, Виктор Астафьев: 25, Иван Тургенев: 25, Лев Пучков: 25, Даниил Гранин: 24, Дарья Вознесенская: 24, Фёдор Крюков: 24, Андрей Лазарчук: 23, Варлам Шаламов: 23, Василий Аксенов: 23, Федор Достоевский: 23, Владимир Тендряков: 22, Евгений Гуляковский: 22, Леонид Андреев: 22, Леонид Словин: 22, Марк Алданов: 22, Сергей Абрамов: 22, Юрий Рытхэу: 22, Антон Чехов: 21, Валентин Пикуль: 21, Виктор Пелевин: 20, Владимир Лосев: 20, Вячеслав Пьецух: 20, Сергей Сергеев-Ценский: 20, Татьяна Форш: 20, Анна Гурова: 19, Василий Песков: 19, Вениамин Каверин: 19, Виктор Дьяков: 19, Владимир Солоухин: 19, Эдуард Хруцкий: 19, Валерий Рощин: 18, Владимир Набоков: 18, Константин Мзареулов: 18, Николай Шпанов: 18, Сергей Баруздин: 18, Александр Бестужев-Марлинский: 17, Алексей Лосев: 17, Виль Липатов: 17, Владимир Гиляровский: 17, Владимир Ильин: 17, Константин Паустовский: 17, Иван Панаев: 17, Мария Семенова: 17, Эдуард Тополь: 17, Валентин Катаев: 16, Василий Ардаматский: 16, Иван Бунин: 16, Аркадий Гайдар: 16, Анатолий Приставкин: 15, Аркадий Адамов: 15, Борис Можаев: 15, Лев Кассиль: 15, Александр Зиновьев: 14, Геннадий Гор: 14, Лев Троцкий: 14, Лев Гумилёв: 14, Лев Толстой: 14, Наталья Мазуркевич: 14, Юрий Поляков: 14, Владимир Сорокин: 13, Михаил Салтыков-Щедрин: 13, Михаил Литов: 13, Олесь Гончар: 13, Еремей Парнов: 13, Александр Куприн: 12, Андрей Платонов: 12, Аристарх Нилин: 12, Виктор Некрасов: 12, Пётр Проскурин: 12, Сергей Довлатов: 12, Чингиз Айтматов: 12, Юрий Трифонов: 12, Юрий Бондарев: 12, Анатолий Безуглов: 11, Борис Житков: 11, Владимир Короленко: 11, Константин Циолковский: 11, Людмила Улицкая: 11, Людмила Петрушевская: 11, Майя Малиновская: 11, Мераб Мамардашвили: 11, Михаил Успенский: 11, Михаил Зуев-Ордынец: 11, Михаил Веллер: 11, Сергей Булгаков: 11, Феликс Разумовский: 11, Юз Алешковский: 11, Александр Чаковский: 10, Алексей Писемский: 10, Вилис Лацис: 10, Владимир Обручев: 10, Владимир Войнович: 10, Владимир Топилин: 10, Вячеслав Шишков: 10, Евгений Гришковец: 10, Илья Эренбург: 10, Константин Симонов: 10, Николай Гарин-Михайловский: 10, Николай Карамзин: 10, Михаил Пришвин: 10, Юрий Герман: 10, Андрей Битов: 9, Борис Полевой: 9, Дмитрий Мережковский: 9, Юлий Дубов: 9, Владимир Одоевский: 8, Григорий Семух: 8, Дмитрий Верищагин: 8, Елена Шилкова: 8, Захар Прилепин: 8, Константин Федин: 8, Михаил Шолохов: 8, Михаил Булгаков: 8, Михаил Задорнов: 8, Михаил Погодин: 8, Николай Гоголь: 8, Николай Успенский: 8, Николай Прокудин: 8, Фёдор Абрамов: 8, Александр Герцен: 7, Александр Солженицын: 7, Анатолий Гладилин: 7, Аркадий

Аверченко: 7, Валентин Распутин: 7, Валентин Леженда: 7, Иван Лажечников: 7, Мариэтта Шагинян: 7, Михаил Загоскин: 7, Юрий Власов: 7, Владимир Соллогуб: 7, Михаил Шторм: 6, Нодар Думбадзе: 6, Овидий Горчаков: 6, Павел Загребельный: 6, Павел Бажов: 6, Сергей Шемякин: 6, Леонид Юзефович: 5, Дмитрий Фурманов: 5.

Важно подчеркнуть, что среди переводных авторов большинство «известных писателей» также оказалось хорошо определяемым:

Барбара Картленд: 342, Роберт Стайн: 208, Эрл Гарднер: 138, Жорж Сименон: 134, Джеймс Чейз: 129, Картер Браун: 120, Агата Кристи: 115, Стивен Кинг: 90, Рекс Стаут: 89, Андрэ Нортон: 85, Пол Андерсон: 83, Жюльетта Бенцони: 80, Александр Дюма: 79, Джон Карр: 67, Дин Кунц: 65, Жюль Верн: 63, Нам Хи Сон: 63, Эд Макбейн: 61, Эдгар Уоллес: 59, Сандра Браун: 58, Артур Конан Дойль: 57, Эдгар Берроуз: 57, Клиффорд Саймак: 56, Терри Пратчетт: 55, Альфред Ван Вогт: 51, Гарри Гаррисон: 51, Иоанна Хмелевская: 51, Айзек Азимов: 49, Оноре де Бальзак: 47, Пэлем Вудхауз: 45, Ребекка Уинтерз: 45, Луи Буссенар: 44, Филип Дик: 44, Дик Фрэнсис: 42, Филип Фармер: 41, Эллери Куин: 41, Глен Кук: 40, Майн Рид: 39, Саймон Грин: 39, Уилбур Смит: 39, Жорж Санд: 38, Алан Фостер: 36, Роберт Хайнлайн: 36, Стефани Лоуренс: 36, Генри Хаггард: 35, Гилберт Честертон: 35, Гарри Тертлдав: 34, Кристофер Сташеф: 34, Майкл Муркок: 34, Роджер Желязны: 34, Эрнст Гофман: 34, Дональд Уэстлейк: 33, Микки Спиллейн: 32, Рафаэль Сабатини: 31, Росс МакДональд: 30, Редьярд Киплинг: 29, Терри Гудкайнд: 28, Джон Гришэм: 27, Джон Кризи: 27, Роберт МакКаммон: 27, Харуки Мураками: 27, Энн Райс: 27, Джек Лондон: 26, Питер Джеймс: 26, Эмиль Золя: 26, Карл Юнг: 25, Питер Чейни: 25, Роберт Ладлэм: 25, Артур Кларк: 23, Астрид Линдгрен: 23, Джон Ле Карре: 23, Зигмунд Фрейд: 22, Шолом-Алейхем: 22, Айрис Мердок: 21, Раймонд Чэндлер: 21, Рэй Брэдбери: 21, Уильям Фолкнер: 21, Франсуаза Саган: 21, Эмилио Сальгари: 21, Эрик ван Ластбадер: 21, Ю Несбё: 21, Алистер Маклин: 20, Альфонс Доде: 20, Герберт Уэллс: 20, Говард Лавкрафт: 20, Джеральд Даррелл: 20, Миранда Ли: 20, О Генри: 20, Урсула Ле Гуин: 20, Артуро Перес-Реверте: 19, Грэм Грин: 19, Фрэнк Херберт: 19, Фрэнсис Фицджеральд: 19, Анджей Збых: 18, Анри Труайя: 18, Брайан Ламли: 18, Джон Голсуорси: 18, Джон Уиндем: 18, Джордж Мартин: 18, Ирвин Шоу: 18, Герман Гессе: 17, Дафна дю Морье: 17, Джакомо Казанова: 17, Роберт Асприн: 17, Сомерсет Моэм: 17, Танит Ли: 17, Томас Манн: 17, Боб Шоу: 16, Джеймс Герберт: 16, Джейн Остин: 16, Дэн Симмонс: 16, Курт Воннегут: 16, Фрэнсис Брет Гарт: 16, Ромен Гари: 15, Франк Тилье: 15, Лион Фейхтвангер: 15, Эдгар По: 15, Ян Флеминг: 15, Богомил Райнов: 14, Милорад Павич: 14, Роберт Силверберг: 14, Франц Кафка: 14, Фред Сейберхэген: 14, Чак Паланик: 14, Вильгельм Гауф: 13, Габриэль Маркес: 13, Джеймс Купер: 13, Джером Клапка Джером: 13, Жан Бодрийяр: 13, Карлос Кастанеда: 13, Ричард Бах: 13, Сельма Лагерлеф: 13, Фредерик Пол: 13, Эльфрида Елинек: 13, Эрнест Хемингуэй: 13, Вальтер Скотт: 12, Джон Апдайк: 12, Жоржи Амаду: 12, Ивлин Во: 12, Кальман Миксат: 12, Лайон Спрэг Де Камп: 12, Фридрих Ницше: 12, Хорхе Луис Борхес: 12, Чарлз Уэбстер Ледбитер: 12, Эрнест Сетон-Томпсон:

12, Барбара Хэмбли: 12, Анатоль Франс: 11, Андрей Гуляшки: 11, Артур Хейли: 11, Джоан Роулинг: 11, Джон Бойнтон Пристли: 11, Марсель Пруст: 11, Милан Кундера: 11, Питер Страуб: 11, Платон: 11, Поль Анри Феваль: 11, Синклер Льюис: 11, Теодор Драйзер: 11, Ахмед Салман Рушди: 10, Джон Кутзее: 10, Джон Стейнбек: 10, Кобо Абэ: 10, Колин Уилсон: 10, Тур Хейердал: 10, Хулио Кортасар: 10, Эдмон Лепеллетье: 10, Леопольд фон Захер-Мазох: 9, Роберт Ханс ван Гулик: 9, Франсуа Мориак: 9, Фредерик Бегбедер: 9, Эммануил Сведенборг: 9, Эрве Базен: 9, Эрих Мария Ремарк: 9, Вашингтон Ирвинг: 8, Генри Миллер: 8, Герман Мелвилл: 8, Джозефина Тэй: 8, Рю Мураками: 8, Себастьян Жапризо: 8, Челси Куинн Ярбро: 8, Эптон Синклер: 8, Андре Жид: 7, Виктор Гюго: 7, Виктор Каннинг: 7, Ги де Мопассан: 7, Даниэль Дефо: 7, Джон Рональд Руэл Толкин: 7, Маркиз де Сад: 7, Бернар Клавель: 6, Борис Виан: 6, Джон О'Хара: 6, Джордж Оруэлл: 6, Марк Туллий Цицерон: 6, Пер Лагерквист: 6, Стефан Цвейг: 6, Франсис Карсак: 6, Чарльз Сноу: 6, Эммануэль Арсан: 6, Ярослав Гашек: 6, Ален Роб-Грийе: 5, Джон Фаулз: 5, Джон Чивер: 5, Марк Твен: 5.

Этот срез авторов позволяет оценить полноту базиса эталонов. Отметим, что перечисленные авторы составляют около 10% правильного корпуса. Следовательно, остальные авторы, тексты которых распознаются безошибочно, хотя и менее популярны, все же не являются «литературным шумом», в котором могут потеряться «известные авторы». Поэтому базис из эталонов не является автоматическим фильтром мастерства. Он важен как наиболее полный набор распределений частот триграмм, позволяющий идентифицировать тексты писателей. Строго говоря, этот базис показывает лишь то, что существует большое число литературных произведений, написанных каждым из авторов в персональном статистически одинаковом стиле. Если использовать построенный базис для повторной идентификации исключенных произведений, то оказывается, что только для 10% этих текстов автор определяется верно. Далее, если автор исключенного текста определен по базису идеального корпуса неверно, но таких текстов у этого автора более двух, их можно попробовать объединить в эталон «автор – 2», если предположить, что в этих произведениях автор реализует другую черту своего писательского мастерства. Такой подход улучшил распознавание только 15% ранее исключенных текстов. Следовательно, произведения, не вошедшие в идеальный корпус, в большинстве случаев (порядка 85%), составляют естественную ошибку метода триграмм. В этой связи важно подчеркнуть, что метод [9] коррекции ошибки позволяет снизить ее в два раза, что показывает более высокое качество этого алгоритма по сравнению с использованием обученной системы.

5. Анализ структуры ошибок для метода триграмм

Построение идеального корпуса позволяет по-новому взглянуть на ошибочно определяемые тексты. Сравним свойства полного и идеального корпусов.

Основное отличие состоит в том, что из полного корпуса были частично удалены относительно короткие тексты – менее 100 тыс. знаков. В основном это относится к текстам длиной от 30 тыс. до 50 тыс. знаков. Для таких произведений частоты триграмм определены недостаточно достоверно, так что применительно к задаче построения авторских эталонов желательно иметь более длинные тексты. На рис. 7 (а, б) показаны распределения корпусов по длине текстов.

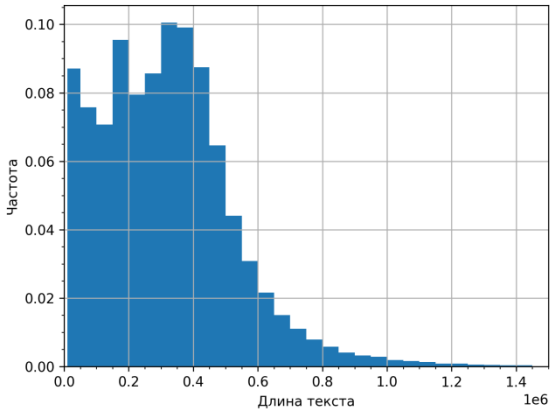

Рис. 7 (а). Полный корпус

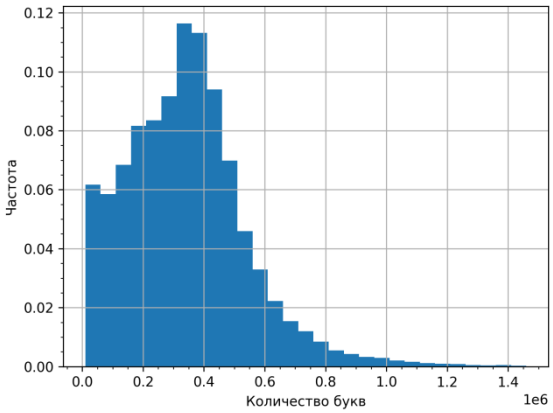

Рис. 7 (б). Идеальный корпус

Как выяснилось, основная часть удаленных текстов относится к жанру «сентиментальный роман». Если в полном корпусе этот жанр занимает первую строчку с долей 0,25, оставляя фантастику на втором месте с долей 0,18, то в идеальном корпусе доля фантастики равна 0,26, а доля любовных романов снижена до 0,15.

В идеальном корпусе в два раза сократилось присутствие классической прозы – с 0,06 до 0,03. Об этом стоит поговорить подробнее. Рассмотрим список русских классиков, имеющих «мировую известность». Это (в алфавитном порядке) Булгаков, Бунин, Гоголь, Горький, Достоевский, Лесков, Набоков, Паустовский, Пришвин, Пушкин, Салтыков-Щедрин, Толстой А., Толстой Л., Тургенев, Шолохов, Чехов. Тексты этих авторов в полном корпусе были идентифицированы с весьма большой ошибкой – на уровне 0,33. Однако если рассмотреть подкорпус только этих классиков, то ошибка распознавания снижается до 0,13 – однако, не до нуля. Возможно, это связано с тем, что эталоны классиков расположены ближе к центру корпуса, чем к периферии – примерно на расстояниях от 0,19 до 0,29, что составляет область второго дециля расстояний. Лишь Пушкин находится относительно далеко от центра корпуса – на расстоянии 0,34. Следовательно, классики используют широкие возможности лексикона в целом, и лишь их индивидуальное мастерство отделяет классические произведения от сериальных авторов. В то же время наличие идеального подкорпуса классиков может быть истолковано как результат исключения второй ипостаси этих авторов.

Процесс удаления ошибочно распознаваемых текстов привел к изменению распределений расстояний между текстами и авторскими эталонами. На рис. 8

показаны распределения расстояний «свой-чужой» для полного и идеального корпусов. Видно, что в результате фильтрации были удалены тексты, которые находились на достаточно большом расстоянии от авторского эталона. Их можно трактовать как произведения, написанные в иной манере и «как бы другим автором». После фильтрации распределение расстояний до своего автора сдвинулось влево в сторону уменьшения расстояний, тогда как распределения расстояний до чужого эталона практически не поменялись. Следовательно, были удалены в основном нетипичные произведения того или иного автора. Фильтрация показала, что ошибка идентификации связана не со статистической неопределенностью в силу большого числа эталонов, а с тем, что авторы иногда могут писать в измененном стиле, что совершенно не запрещено.

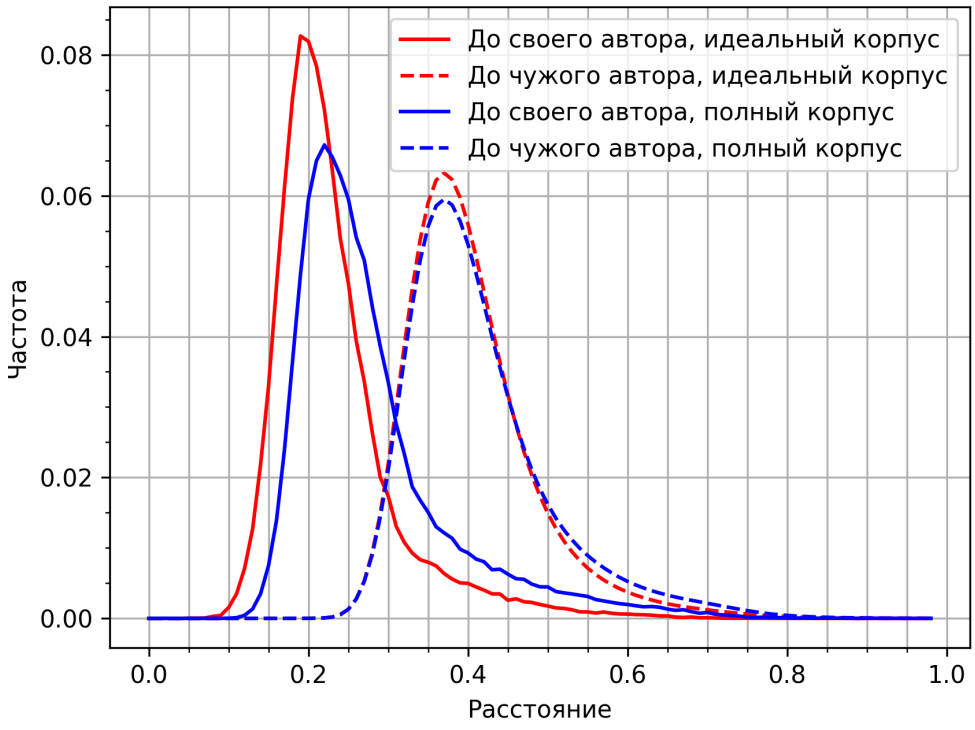

Рис. 8. Распределения расстояний между текстами и эталонами

При формировании идеального корпуса были удалены 2416 авторов и 31857 текстов. Для них ошибка идентификации имеет следующую структуру. Примерно 40% текстов перепутывается с эталонами авторов, которых можно считать ближайшими соседями (первый – десятый) правильного автора. Примерно по 20% приходится на соседей порядков от 10 до 50 и от 100 до 1000. Еще примерно по 5% приходится на соседей порядков от 50 до 100 и на совсем дальних соседей порядка более 1000. На рис. 9 приведено распределение расстояний первых ближайших соседей для эталонов авторов полного корпуса.

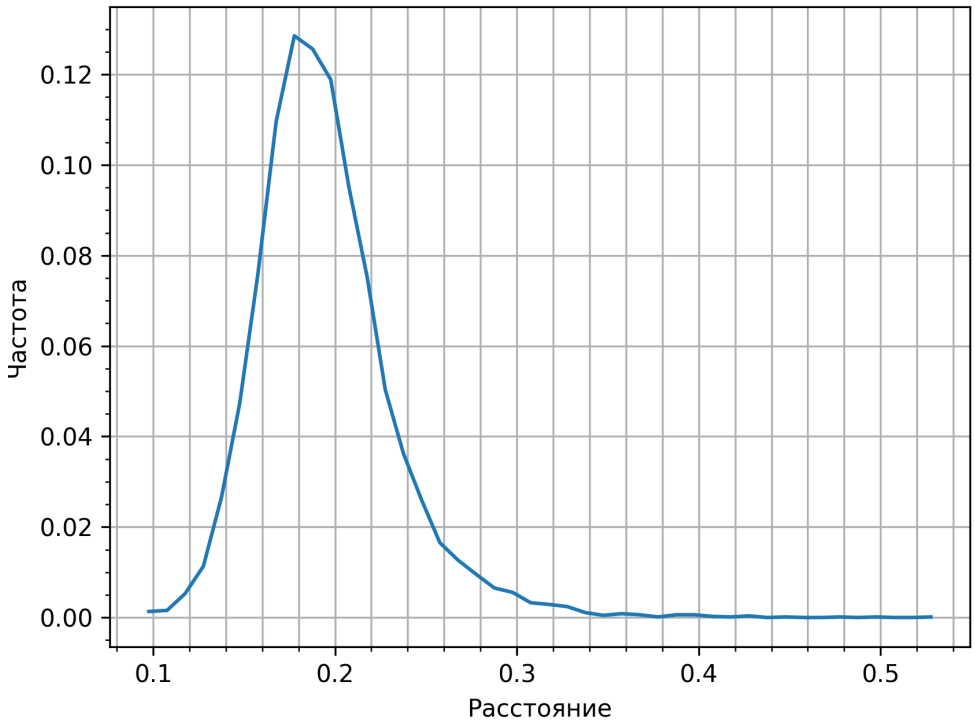

Рис. 9. Распределение расстояний между первыми ближайшими эталонами

Эти наблюдения и позволили сформулировать гипотезу коррекции ошибок, основанную на идее о выдерживании автором своего эталонного распределения. Те авторы, которые перепутываются с ближайшими соседями, производят собственно ошибку метода. Расстояния между эталонами этой группы авторов меньше, чем точность, с которой определяются эмпирические частоты, поэтому ошибка носит неустранимый случайный характер. В группе авторов, для которых перепутывание происходит с соседями порядков от 10 до 100, преобладает «самиздат» и переводная беллетристика, т.е. ошибочно идентифицируются тексты не вполне профессиональных писателей. И, наконец, тексты, которые перепутываются с эталонами дальних соседей, естественно считать написанными в манере, значительно отличающейся от текстов эталона правильного автора. Для этих последних, доля которых в общем числе ошибок составляет примерно половину, и предложен метод подтверждения авторства при делении текста на фрагменты.

Отметим, что среди 2,5 тыс. исключенных авторов не так много оказалось «известных писателей» – примерно 2%. Эти писатели следующие.

Российские авторы (15 писателей): Иван Гончаров, Евгений Замятин, Николай Рерих, Борис Лавренев, Нина Берберова, Анна Старобинец, Татьяна Толстая, Мария Арбатова, Михаил Зощенко, Николай Чернышевский, Герман Садулаев, Юрий Олеша, Николай Носов, Дмитрий Нагишкин, Всеволод Гаршин.

Зарубежные авторы (30 писателей): Гор Видал, Мартин Хайдеггер, Питер Хёг, Вольтер, Адольфо Биой Касарес, Мигель де Сервантес, Питер Устинов,

Теофиль Готье, Фридрих Дюренматт, Итало Кальвино, Жан-Поль Сартр, Вирджиния Вульф, Проспер Мериме, Джеймс Олдридж, Питер Бенчли, Болеслав Прус, Дэн Браун, Уильям Стайрон, Патрик Квентин, Фарли Моуэт, Кен Кизи, Морис Дрюон, Альфред де Мюссе, Этель Войнич, Оскар Уайлд, Хаймито фон Додерер, Луи Арагон, Карел Чапек, Норман Льюис, Рюноксэ Акутагава.

Таким образом, статистический метод идентификации автора показал, что большинство профессиональных писателей пишут, сохраняя не только узнаваемый авторский стиль, но и воспроизводя функцию распределения буквосочетаний. Именно последнее обстоятельство позволяет использовать методы машинного обучения для автоматической классификации текстов, поскольку оно базируется на объективном подсчете формальных символов. Предъявленные исключения при построении идеального корпуса аргументированно относятся только к русскоязычным писателям, так как при переводах могут быть искажения авторского стиля разными переводчиками. Тем не менее, для коррекции модели было бы полезно понять причины того, почему перечисленные авторы не сохраняют статистические свойства своего творчества. Однако это направление выходит за рамки собственно математической модели.

Заключение

В данной работе мы рассмотрели два аспекта, влияющие на ошибку распознавания автора текста. Эти аспекты – близость авторского эталона к среднему эталону корпуса, а также ширина среднего носителя текстов автора. Выяснилось, что примерно 80% текстов имеют монотонную зависимость ошибки от данных параметров. Интересно отметить, что точность идентификации автора текста методом триграмм также порядка 80%, но это, видимо, обычное совпадение. Первый и последний децили рассматриваемых зависимостей состоят из авторов, которые имеют различный уровень мастерства, но, видимо, из-за того, что «крайности смыкаются», именно в этих кластерах нет устойчивого характера ошибки как функции параметров текста.

Выявлены базовые тенденции, связанные с зависимостью ошибки распознавания автора текста от параметров эталона автора – близости к центру корпуса и ширины носителя. Эти параметры демонстрируют противоположные тенденции, а именно, чем ближе эталон находится к центру корпуса, тем хуже распознается текст такого автора, но чем шире средний носитель текста, тем меньше ошибка. Интересно же то, что с приближением эталона к центру корпуса увеличивается ширина его носителя. Следовательно, широкий носитель и некоторая удаленность от центра могут характеризовать степень писательского мастерства автора. Хотя, конечно, это не общий писательский закон, а лишь наблюдаемая в среднем тенденция.

Весьма важным результатом проведенной «работы над ошибками» является возможность их полного исключения. В результате был построен

идеальный корпус, каждый текст которого идентифицируется правильно. В определенном смысле идеальный корпус представляет собой статистически настраиваемый базис из эталонов, который получается в ходе машинного обучения. Однако далее мы можем наблюдать ту самую картину, которую обычно стараются не афишировать: применение обученной машины вне зоны обучения дает неудовлетворительные результаты. В нашем случае с помощью идеального базиса удалось уточнить всего лишь 10% текстов, не вошедших в этот корпус. Остальные тексты по-прежнему распознавались неверно. Это связано с тем, что если базис из распределений построен по стационарным данным, а таковыми всегда являются данные конечных выборок, то он плохо работает в нестационарной области, когда данные будут поступать из других распределений.

Разделы 2, 3 и 4 написаны М.Ю. Кислицыной, остальные разделы написаны совместно.

Благодарности
Работа выполнена при поддержке гранта РНФ, проект № 23-71-10055.

Литература

1. Кислицына М.Ю., Орлов Ю.Н. Статистический анализ полного корпуса художественной литературы на русском языке и распознавание автора // Препринты ИПМ им. Келдыша. 2024. № 17. С. 1-24.
2. Рассел С., Норвиг П. Искусственный интеллект. Современный подход. – М.: Вильямс. 2007. – 1480 с.
3. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. – М.: Вильямс, 2011. – 528 с.
4. Hovy E., Lavid Ju. Towards a science of corpus annotation // International Journal of Translation, 2010. V. 22. No 1. P. 1-25.
5. Шевелев О.Г. Методы автоматической классификации текстов на естественном языке: Учебное пособие. – Томск: ТМЛ-Пресс, 2007. – 144 с.
6. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 326 с.
7. Резанова З.И., Романов А.С., Мещеряков Р.В. О выборе признаков текста, релевантных в автороведческой экспертной деятельности // Вестник Томского государственного университета. Филология. – 2013. –Т. 26. № 6. – С. 38-52.
8. Воронина М.Ю., Орлов Ю.Н. Определение автора текста методом сегментации // Компьютерные исследования и моделирование. 2022. Т. 14. № 5. С. 1199-1210.

9. Воронина М.Ю., Кислицын А.А., Орлов Ю.Н. Алгоритм коррекции метода биграмм в задаче идентификации автора текста // Математическое моделирование. 2022. Т. 34. № 9. С. 3-20.
10. Корпус текстов.
https://github.com/akislitsyn/textcorpus (последняя дата обращения 03.03.2024)