# Раздел III. Математическое и программное обеспечение

УДК 004.931 DOI 10.23683/2311-3103-2018-8-104-114

**Д.А. Бирин, С.Ю. Мельников, В.А. Пересыпкин, И.А. Писарев, Н.Н. Цопкало**

## ОБ ЭФФЕКТИВНОСТИ СРЕДСТВ КОРРЕКЦИИ ИСКАЖЕННЫХ ТЕКСТОВ В ЗАВИСИМОСТИ ОТ ХАРАКТЕРА ИСКАЖЕНИЙ

*Анализируются возможности четырех программных средств автоматической коррекции текстов (Яндекс.Спеллер, Afterscan, Bing Spell Check, Texterra) для коррекции искаженных текстов. Описаны искажения текстов, возникающие при вводе текста с клавиатуры и работе систем распознавания. Для перечисленных программных средств приводятся данные экспериментов по точности коррекции искаженных текстов, полученных как при клавиатурном вводе текста, так и на выходе систем оптического распознавания текста при обработке изображений плохого качества и распознавания речи в условиях шумов. Для моделирования искажений, вносимых системами распознавания, предложена двухэтапная модель случайных искажений текстов. На первом этапе (словарные искажения с заданной вероятностью) искажаемое слово в тексте заменяется на случайное словарное слово, удаленное от искажаемого на расстояние Левенштейна 1 или 2. Выбор заменяющего слова производится по равновероятной схеме. На втором этапе (символьные искажения с заданной вероятностью) искажаемый знак текста с вероятностью 1/3 либо удаляется, либо перед ним осуществляется вставка случайного символа, либо искажаемый знак заменяется на случайный символ алфавита. Выбор случайного символа производится по равновероятной схеме. Полученные таким образом искаженные тексты исправляются с помощью выбранных программных средств и подсчитывается процент истинных слов в скорректированном тексте. Полученные данные усредняются по набору текстов. Приводятся результаты экспериментов с оценкой точности коррекции в следующей области параметров: вероятности словарного искажения изменяются от 0.01 до 0.9 и вероятности символьного искажения изменяются от 0.01 до 0.5. Полученные результаты показывают, что Яндекс.Спеллер, Bing Spell Check и Texterra обеспечивают хорошее качество коррекции искажений, возникающих при клавиатурном вводе. Для коррекции искажений, вносимых системами распознавания, перечисленные программные средства неэффективны.*

*Искаженные тексты; случайные искажения; автоматическая коррекция; постобработка.*

**D.A. Birin, S.Yu. Melnikov, V.A. Peresypkin, I.A. Pisarev, N.N. Copkalo**

## ON THE EFFICIENCY OF THE NOISY TEXT CORRECTION SOFTWARE DEPENDING ON THE DISTORTION TYPE

*The capabilities of four automatic text correction software (Yandex.Speller, Afterscan, Bing Spell Check, Texterra) for noisy texts correction are analyzed. The distortions of texts that occur while typing text on the keyboard and recognition systems working are described. Experimental data on the accuracy of the correction of distorted texts obtained both by typing and as the output of real OCR systems processing low-quality images and ASR systems in a noisy environment are presented. To simulate the distortions caused by the recognition systems, a two-stage model of random text distortions is proposed. At the first stage (word distortions with a given probability) the distorted word in the text is replaced with a random dictionary word with Levenshtein distance 1 or 2. The replacement word is chosen according to the uniform distribution. At the second stage (character distortions with a given probability) the distorted character is removed with a probabil-*

*ity of 1/3, or a random character is inserted before it with a probability of 1/3, or it is replaced with a random alphabet character with a probability of 1/3. The replacement character is chosen according to the uniform distribution. The distorted texts obtained in this way are corrected using the Yandex.Speller and Bing Spell Check software and the percentage of true words in the corrected text is calculated. The data are averaged over a set of texts. The results of experiments with an estimation of the correction accuracy in the following parameter range are given: the probabilities of word distortion vary from 0 to 0.9 and the probabilities of symbol distortion vary from 0 to 0.5. The results show that Yandex.Speller, Bing Spell Check and Texterra provide good quality of the correction of distortions that occur while typing. This software are ineffective for correcting distortions caused by the recognition systems.*

*Noisy texts; random distortions; automatic correction; post-processing.*

**Введение.** Одним из основных факторов, существенно затрудняющих понимание, перевод и анализ текста, являются искажения в нем, возникающие в результате опечаток при клавиатурном вводе, ошибок при автоматическом распознавании речи и оптическом распознавании изображений текстов. Такие тексты содержат ошибочные символы, слова и словосочетания, и при большом количестве искажений весьма тяжелы для восприятия [1]. В [2] приведена классификация возможных типов ошибок в текстах, возникающих при наборе (в том числе, при переписке в социальных сетях), при оптическом распознавании и распознавании речи, при машинном переводе. Тексты с такими ошибками называют искаженными, или зашумленными («noisy texts»).

В работах [3–4] и [5–8] представлены подходы к «post-ASR» и «post-OCR» обработке, то есть методы коррекции искаженных текстов, полученных в результате распознавания речи и оптического распознавания.

В [5] отмечается, что качество полученных после использования систем оптического распознавания исторических текстов (рассматривались, в частности, материалы австралийской газеты «Sydney Morning Herald» за 1842–1954 гг.) недостаточно для эффективной работы с ними с помощью поисковых инструментов. Описана архитектура системы постобработки для автоматической коррекции исторических текстов, которые были получены в результате оптического распознавания с недостаточным качеством. Ошибки распознавания влияют на точность подсистем автоматического анализа текста: определение границ предложений, токенизация, разметка по частям речи [15], выделение названий и фамилий [7]. В [8] описана технология обработки искаженных текстов, полученных с помощью системы оптического распознавания для письменности Деванагари. Используется матрица ошибок символов, построенная на большом корпусе текстов на хинди. С помощью матрицы ошибок порождаются слова – кандидаты на замену ошибочных.

Для исследования влияния зашумленности текста на работу систем машинного перевода в [9] предложено разработать параллельные корпуса с искусственно внесенными искажениями. Рассматриваются следующие типы искажений: изменение регистра отдельных букв, замена слов на фонетически схожие, выброс одной или нескольких букв в слове для его сокращения, выброс слов из предложения для его сокращения, слияние соседних слов в предложении.

В работах [10–15] изучаются искаженные тексты естественной природы, то есть созданные человеком. Искажения в таких текстах вносятся самими пользователями при написании для скорости письма или по неграмотности. В [10] оценивается влияние вносимых пользователями искажений на точность работы автоматических систем разметки текста по частям речи и систем выделения мнений. В [11–15] предложены способы автоматической коррекции (нормализации) искаженных текстов сообщений в социальных сетях для английского, малайского, китайского, уйгурского и японского языков.

В [3] и [16] описаны технологии коррекции искаженных текстов с помощью онлайн средств от Microsoft и Google.

Для оценки возможности исправления искажений в тексте проведены исследования нескольких доступных программных средств коррекции искаженных текстов, предназначенных для коррекции ошибок ручного ввода текста с клавиатуры, оценена их эффективность при коррекции текстов, образующихся на выходе систем распознавания и при коррекции текстов со случайными искажениями.

**Программные средства коррекции искаженных текстов**

1. *Яндекс.Спеллер.* Сервис проверки правописания Яндекс.Спеллер предлагает веб-разработчикам использовать возможность проверки орфографии на страницах своих сайтов [17]. Спеллер анализирует слова, основываясь на правилах орфографии и лексике современного языка. Используется орфографический словарь, содержащий правильные написания большинства наиболее употребимых слов.

2. *AfterScan.* AfterScan – программа от разработчика InteLife Solutions, предназначенная для автоматического корректирования текстов [18]. К ее основным возможностям относятся: проверка орфографии, автоматическое исправление ошибок распознавания/ручного ввода, приведение к типографским нормам. Программа имеет интерфейс, подобный другим текстовым редакторам (WordPad, Microsoft Word и т.д.).

3. *Texterra.* Технология автоматического построения онтологий и семантического анализа текста разработана в Институте системного программирования РАН [19]. Состоит из трех модулей: лингвистического анализа, базы знаний, извлечения информации. Модуль лингвистического анализа выполняет морфологический анализ слов, синтаксический анализ предложений, поиск кореферентных цепочек слов, орфографическую коррекцию. Модуль базы знаний использует в качестве источника данных Википедию и Викиданные, позволяет производить подсчет семантической близости между понятиями. Модуль извлечения информации производит привязку фрагментов текста к понятиям базы знаний и извлечение основных понятий текста.

4. *Bing Spell Check.* Средство, разработанное Microsoft, позволяет выполнить контекстную проверку грамматики и орфографии [20]. Проверка орфографии основана на использовании корпуса веб-поиска и документов. Поддерживается две модели исправления: Proof и Spell. Первая корректирует текст максимально качественно, она и использовалась в дальнейших исследованиях.

Результаты сравнительного анализа выбранных систем коррекции сведены в табл. 1. В плане функциональности можно выделить Яндекс.Спеллер и Bing Spell Check. В плане удобства пользовательского интерфейса предпочтительнее Яндекс.Спеллер и Texterra, в частности, имеется подсветка неправильно написанных слов. API для использования сервисов сторонними приложениями присутствует у всех систем, кроме AfterScan, и представляет собой web API, взаимодействие с которым производится путем POST или GET запросов. Открытой для изучения алгоритмов её функционирования является система Texterra, остальные системы являются коммерческими и их внутреннее описание недоступно.

Полностью бесплатными системами без ограничения функционала являются Яндекс.Спеллер и Texterra. AfterScan и Bing Spell Check являются частично бесплатными, имея бесплатные версии или ограничивая бесплатный доступ временными рамками.

Таблица 1

**Сравнительный анализ программных средств коррекции**

| Критерий | Яндекс.Спеллер | AfterScan | Texterra | Bing Spell Check |
|---|---|---|---|---|
| Функциональность, 1-5 | 5 | 3 | 4 | 5 |
| Удобство интерфейса, 1-5 | 5 | 4 | 5 | 3 |
| Наличие API, да/нет | да | нет | да | да |
| Открытость, да/нет | нет | нет | да | нет |
| Бесплатность, да/частично/нет | да | частично | да | частично |

**Искажения текстов, возникающие при работе систем распознавания и вводе текста с клавиатуры.** Качество распознавания речи зависит от нескольких факторов: качество записи, наличие шумов, особенности речи. По результатам проведенного анализа ошибок распознавания выделены наиболее характерные типы: - неправильная форма слова; - замена слова на похожее по звучанию; - замена нескольких слов на одно; - замена одного слова несколькими; - пропуск слов; - вставка или удаление коротких слов (предлогов и союзов). В результате распознавания получается текст, имеющий некоторые искажения и состоящий, в основном, из словарных слов, в том числе и в местах искажений [21].

При распознавании изображений текста возникают ошибки, при которых часть символов заменяется на близкие по написанию. В результате встроенной в OCR системы словарной корректировки повышается качество распознавания, а полученный текст содержит, в основном, словарные слова. В некоторых случаях неправильно распознанные символы позволяют получить слово, корректное по написанию, но отличное от использованного в исходном тексте. Так же характерными для распознанного текста ошибками являются: два или более слов на изображении распознаются как одно слово, часть символов которого совпадает по расположению с истинными словами, и обратный случай, когда вместо одного слова может быть распознано несколько слов меньшей длины [22, 23].

При вводе текста с помощью клавиатуры возможны следующие ошибки: - замена текущего символа на рядом стоящий символ на клавиатуре; - пропуск символа; - вставка рядом стоящего символа на клавиатуре либо перед, либо после истинного символа; - дублирование символа; - изменение порядка двух рядом стоящих символов в слове [24].

**Коррекция текстов с модельными случайными искажениями.** Из перечня рассмотренных выше программных средств автоматической коррекции текстов было выбрано два, Яндекс.Спеллер и Bing Spell Check. По корпусу новостных текстов на английском языке [25] построен словарь объема 230 000 слов. Были случайно отобраны 30 текстов, объем каждого из которых составил от 150 до 750 слов.

Использовалась следующая двухэтапная схема искажений текстов. На первом этапе последовательно просматривались все слова входного текста. С вероятностью $P_2$ для текущего слова принималось решение (независимо от решений, принятых для предыдущих слов), будет ли оно подвергнуто искажению. В случае принятия такого решения текущее слово заменялось на другое, выбранное равновероятно из слов словаря, находящихся на расстоянии Левенштейна $L = 1, 2$ от текущего.

На втором этапе последовательно просматривались все символы входного текста, за исключением тех, которые составляют измененные слова, полученные на первом этапе. С вероятностью $P_1$ для текущего символа принималось решение

(независимо от решений, принятых для предыдущих символов), будет ли он подвергнут искажению. Если принималось решение, что символ подвергается искажению, то с вероятностью 1/3 он удалялся, с вероятностью 1/3 перед ним вставлялся символ из алфавита, выбираемый равновероятно, с вероятностью 1/3 истинный символ заменялся на равновероятно выбранный символ алфавита. Значения $P_1$, $P_2$ и $L$ использовались в качестве параметров.

Полученные таким образом искаженные тексты исправлялись с помощью программных средств Яндекс.Спеллер и Bing Spell Check. В искаженных и скорректированных текстах подсчитывался средний по 30 испытаниям процент $Noisy(P_1,P_2)$, $Yand(P_1,P_2,L)$ и $Bing(P_1,P_2,L)$ верных слов по отношению к числу всех слов исходного неискаженного текста, соответственно.

Результаты проведенных экспериментов представлены на рисунках ниже.

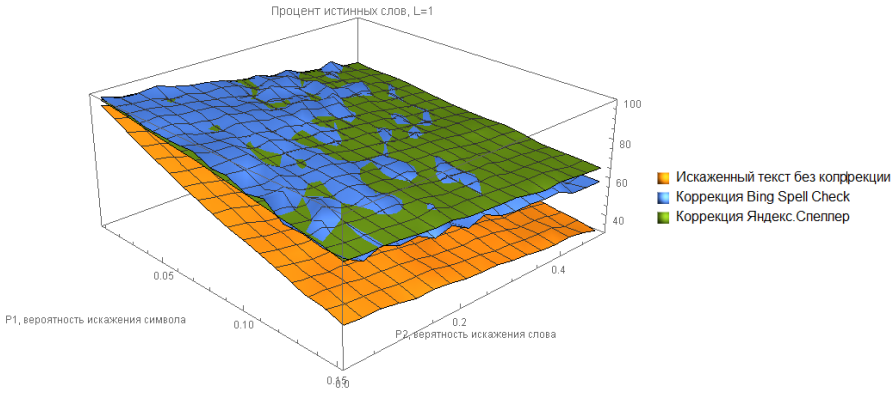


*Рис. 1. Поверхности* $Noisy(P_1,P_2)$, $Yand(P_1,P_2,1)$ *и* $Bing(P_1,P_2,1)$,
$0 < P_1 < 0.15, 0 < P_2 < 0.5$

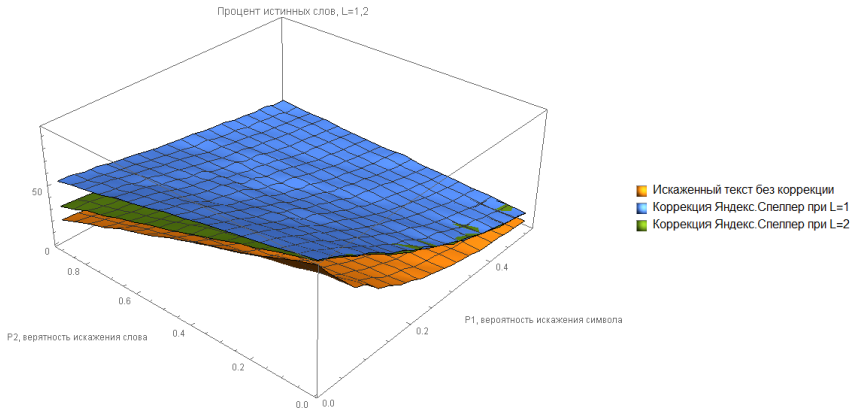


*Рис. 2. Поверхности* $Noisy(P_1,P_2)$, $Yand(P_1,P_2,1)$, $Yand(P_1,P_2,2)$,
$0 < P_1 < 0.5, 0 < P_2 < 0.9$

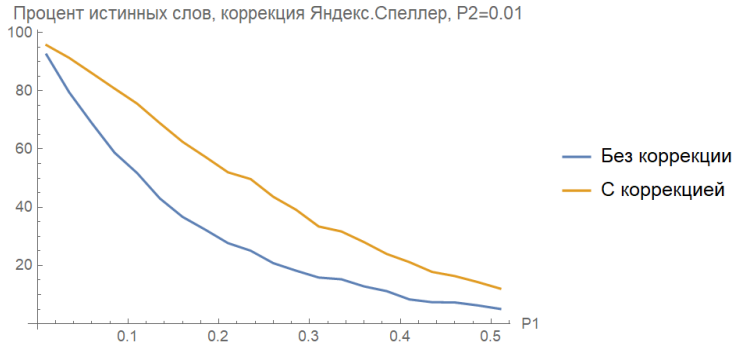


*Рис. 3. Процент истинных слов в зависимости от вероятности* $P_1$ $(P_2 = 0.01)$

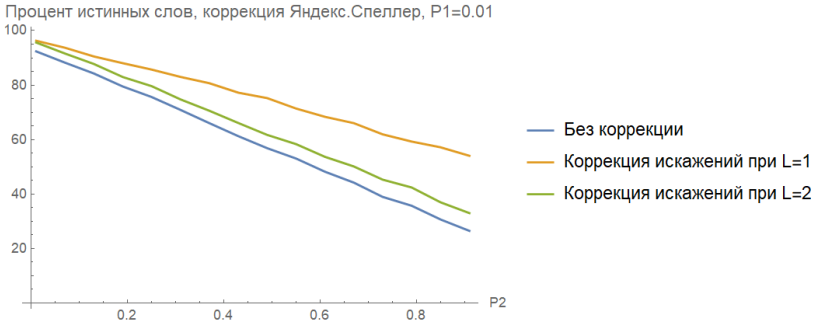


*Рис. 4. Процент истинных слов в зависимости от вероятности* $P_2$ $(P_1 = 0.01)$

Приведенные результаты показывают, что обе системы, Яндекс.Спеллер и Bing Spell Check, обладают близкими характеристиками по точности коррекции текстов с рассматриваемыми характеристиками искажений. Вид поверхностей $Noisy(P_1, P_2)$, $Yand(P_1, P_2, 1)$ и $Bing(P_1, P_2, 1)$, $0 < P_1 < 0.15, 0 < P_2 < 0.5$, приведенный на Рис.1, говорит о том, что тексты с низкими значениями искажений Bing Spell Check корректирует несколько лучше, а при сильных искажениях лучшие результаты у Яндекс.Спеллер. Результаты, приведенные на Рис.2 и Рис.3, показывают, что в диапазоне искажений $0 < P_1 < 0.1, 0 < P_2 < 0.1$ Яндекс.Спеллер обеспечивает повышение процента истинных слов в тексте с уровня приблизительно 50% до уровня приблизительно 70% при $L = 1$ и до уровня приблизительно 60% при $L = 2$. График, приведенный на Рис.4, демонстрирует, что качество коррекции текстов для случая $L = 2$ значительно хуже. Таким образом, рассматриваемые системы неэффективны при коррекции замен слова на другое словарное слово при $L = 2$.

**Коррекция текстов с реальными искажениями.** Для оценки возможности коррекции искаженных текстов, возникающих при ручном вводе оператором текстов с клавиатуры, были использованы 30 текстов объемом до одной страницы (1000–1500 символов), набранных оператором с клавиатуры, результаты коррекции 10 типичных из них приведены в табл. 2. Точность работы каждой из рассмотренных выше программ коррекции отображена в двух столбцах: в первом показан процент истинных слов после коррекции, а во втором – разность между процентами истинных слов в скорректированном и исходном искаженном тексте.

Как видно из представленных в табл. 2 данных, системы коррекции Bing, Яндекс.Спеллер и Texterra, в целом, справляются с задачей коррекции искажений, снижая исходный уровень искажений примерно вдвое. Bing корректирует, в среднем, немного лучше, а Texterra – немного хуже. Программная система AfterScan оказалась явным аутсайдером, практически не выполнив коррекцию искаженных текстов.

Таблица 2

**Результаты коррекции ошибок ввода текста с клавиатуры**

| Текст, % | Яндекс.Спеллер | | AfterScan | | Texterra | | Bing | |
|---|---|---|---|---|---|---|---|---|
| 84,77 | 90,07 | 5,30 | 84,77 | 0,00 | 90,07 | 5,30 | 96,69 | 11,92 |
| 80,31 | 94,15 | 13,84 | 80,62 | 0,31 | 92,92 | 12,61 | 92,00 | 11,69 |
| 77,85 | 95,38 | 17,54 | 78,15 | 0,31 | 88,31 | 10,46 | 98,15 | 20,31 |
| 75,19 | 92,74 | 17,55 | 75,95 | 0,76 | 86,25 | 11,06 | 94,27 | 19,08 |
| 74,72 | 95,51 | 20,79 | 74,72 | 0,00 | 91,01 | 16,29 | 98,31 | 23,60 |
| 70,20 | 88,74 | 18,54 | 70,20 | 0,00 | 85,43 | 15,23 | 95,36 | 25,17 |
| 68,89 | 88,25 | 19,37 | 69,52 | 0,63 | 86,67 | 17,78 | 95,56 | 26,67 |
| 68,34 | 94,96 | 26,62 | 69,78 | 1,44 | 84,17 | 15,83 | 97,84 | 29,50 |
| 64,57 | 89,37 | 24,80 | 65,35 | 0,79 | 83,07 | 18,50 | 92,91 | 28,35 |
| 63,78 | 82,05 | 18,27 | 64,10 | 0,32 | 81,41 | 17,63 | 88,46 | 24,68 |
| Среднее: 72,86 | 91,12 | 18,26 | 73,31 | 0,45 | 86,93 | 14,07 | 94,95 | 22,09 |

Аналогичные исследования были проведены для 30 искаженных текстов объемом около одной страницы, полученных в результате распознавания записей, проведенных в условиях шумов, прочтенных диктором текстов с помощью специализированной системы распознавания речи, и на выходе OCR-системы FineReader ver.12 при распознавании сканированных изображений газет 60-80-х годов прошлого века. Как следует из представленных в табл. 3 результатов коррекции для типичных 10 исследованных текстов, происходит минимальное снижение количества искажений в тексте, а в результате работы программной системы AfterScan количество искаженных слов возрастает.

Таблица 3

**Результаты коррекции ошибок систем распознавания**

| Текст % | Яндекс.Спеллер | | AfterScan | | Texterra | | Bing | |
|---|---|---|---|---|---|---|---|---|
| 91,00 | 94,07 | 3,07 | 91,21 | 0,21 | 93,66 | 2,66 | 95,91 | 4,91 |
| 91,43 | 95,43 | 4,00 | 90,86 | -0,57 | 93,71 | 2,28 | 97,71 | 6,28 |
| 90,46 | 94,15 | 3,69 | 90,77 | 0,31 | 92,62 | 2,16 | 96,31 | 5,85 |
| 82,37 | 85,26 | 2,89 | 82,05 | -0,32 | 86,86 | 4,49 | 88,78 | 6,41 |
| 85,88 | 91,72 | 5,84 | 85,55 | -0,33 | 89,45 | 3,57 | 91,07 | 5,19 |
| 82,33 | 82,33 | 0,00 | 81,27 | -1,06 | 81,98 | -0,35 | 81,98 | -0,35 |
| 83,44 | 82,78 | -0,66 | 80,13 | -3,31 | 82,78 | -0,66 | 82,33 | -1,11 |
| 73,25 | 76,13 | 2,88 | 72,02 | -1,23 | 76,13 | 2,88 | 76,95 | 3,70 |
| 71,48 | 72,13 | 0,65 | 70,49 | -0,98 | 71,48 | 0,00 | 72,79 | 1,31 |
| 78,46 | 79,08 | 0,62 | 73,23 | -5,23 | 78,46 | 0,00 | 79,38 | 0,92 |
| Среднее: 72,86 | 85,31 | 2,30 | 81,76 | -1,25 | 84,71 | 1,70 | 86,32 | 3,31 |

**Основные результаты и выводы.** Применение рассмотренных программных средств автоматической коррекции, обеспечивающих хорошее качество коррекции ошибок клавиатурного ввода, для текстов, получающихся при автоматическом распознавании зашумленной речи, оптическом распознавании изображений текстов низкого качества, и текстов, искаженных программно с использованием случайных словарных замен, является неэффективным. Это делает необходимым разработку новых подходов к коррекции искаженных текстов, основанных на результатах анализа специфики искажений, использовании моделей текстов, учитывающих глубокие контекстные зависимости между словами, и разработке эффективных алгоритмов поиска варианта скорректированного текста.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК


1. *Бирин, Д.А., Мельников С.Ю., Пересыпкин В.А*. Об эффективности средств коррекции искаженных текстов для результатов работы систем распознавания // Суперкомпьютерные технологии (СКТ-2018): Материалы 5-й Всероссийской научно-технической конференции: в 2 т. – Т. 1. – Ростов-на-Дону; Таганрог: Изд-во ЮФУ, 2018. – С. 71-75.
2. *Subramaniam L.V. et al.* A survey of types of text noise and techniques to handle noisy text // Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, July 23-24, 2009, Barcelona, Spain. DOI: 10.1145/1568296.1568315.
3. *Bassil Y., Alwani M.* Post Editing Error Correction Algorithm for Speech Recognition using Bing Spelling Suggestion // International Journal of Advanced Computer Science and Applications. – 2012. – Vol. 3, No.2. – P. 95-101.
4. *Feld M., Momtazi S., Freigang F., Klakow D., Müller C.* Mobile texting: can post-ASR correction solve the issues? An experimental study on gain vs. costs // Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, February 14-17, 2012. – P. 37-40. Lisbon, Portugal. DOI: 10.1145/2166966.2166974.
5. *Evershed J., Fitch K.* Correcting Noisy OCR: Context beats Confusion DATeCH 2014, May 19–20, 2014, Madrid, Spain DOI:10.1145/2595188.2595200.
6. *Lopresti D.P.* Optical character recognition errors and their effects on natural language processing // International Journal on Document Analysis and Recognition (IJDAR). – September 2009. – Vol. 12, Issue 3. – P. 141–151. DOI: 10.1007/s10032-009-0094-8.
7. *Packer T.L., Lutes J.F., Stewart A.P., Embley D.W., Ringger E.K., Seppi K.D., et al.* Extracting person names from diverse and noisy OCR text // Proceedings of the fourth workshop on Analytics for noisy unstructured text data AND '10, 2010. – P. 19-26. DOI 10.1145/1871840.1871845.
8. *Kumar A., Lehal G.S.* Automatic Text Correction for Devanagari OCR // Indian Journal of Science and Technology. – December 2016. – Vol. 9 (45). DOI: 10.17485/ijst/2016/v9i45/106372.
9. *Gadde P., Goutam R., Shah R., Bayyarapu H.S., Subramaniam L.V.* Experiments with artificially generated noise for cleansing noisy text // Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, MOCR AND '11. – P. 4:1-4:8. ACM, 2011.
10. *Dey L., Haque S.K.M.* Studying the effects of noisy text on text mining applications // Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data AND'09. – Barcelona, Spain, 2009. – P. 107-114.
11. *Clark E., Araki K.* Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English // Procedia - Social and Behavioral Sciences 27, December 2011. – P. 2-11. DOI: 10.1016/j.sbspro.2011.10.577.
12. *Saloot M.A., Idris N., Mahmud R.* An architecture for Malay Tweet normalization // Inf. Process. Manag. – 2014. – Vol. 50, No. 5. – P. 621-633, DOI: 10.1016/j.ipm.2014.04.009.
13. *Wang A., Kan M.-Y., Andrade D., Onishi T., Ishikawa K.* Chinese Informal Word Normalization: an Experimental Study // International Joint Conference on Natural Language Processing. – 2013. – P. 127-135. DOI: 10.1007/978-3-319-68612-7_25.
14. *Tursun O., Cakici R.* Noisy Uyghur Text Normalization // Proceedings of the 3rd Workshop on Noisy User-generated Text, pp. 85–93, Copenhagen, Denmark, September 7, 2017. DOI: 10.18653/v1/w17-4412.

15. *Ikeda T., Shindo H., Matsumoto Y.* Japanese Text Normalization with Encoder-Decoder Model // Proceedings of the 2nd Workshop on Noisy User-generated Text. – Osaka, Japan, December 11, 2016. – P. 118-126.
16. *Bassil, Y., Alwani, M.* OCR post-processing error correction algorithm using Google's online spelling suggestion // Journal of Emerging Trends in Computing and Information Sciences. – January 2012. – Vol. 3, No. 1.
17. *Спеллер* – Технологии Яндекса. – URL: https://tech.yandex.ru/speller/ (accessed: 08.11.2018).
18. *AfterScan* – post-OCR text proofing, advanced spell-checking, automatic correction. – URL: http://www.afterscan.com/ru/ (accessed: 08.11.2018).
19. *Турдаков Д. и др.* Texterra: инфраструктура для анализа текстов // Труды Института системного программирования РАН. – 2014. – Т. 26. – Вып. 1. – С. 421-438. DOI: 10.15514/ISPRAS-2014-26(1)-18.
20. *Microsoft* Cognitive Services – API Bing проверки орфографии. – URL: https://azure.microsoft.com/ru-ru/services/cognitive-services/spell-check/ (accessed: 08.11.2018).
21. *Мещеряков Р.В.* Структура систем синтеза и распознавания речи // Известия Томского политехн. ун-та. – 2009. – Т. 315, № 5. – С. 127-132.
22. *Смирнов С.В.* Корректировка ошибок оптического распознавания на основе рейтингоранговой модели текста // Труды СПИИРАН. – 2014. – Вып. 4, № 35. – С. 64-82. DOI: 10.15622/sp.35.5.
23. *Рудаков И.В., Романов А.С.* Распознавание текстового изображения с учетом морфологии слова // Наука и образование: научное издание МГТУ им. Н.Э. Баумана. – 2012. –Вып. 4. – С. 1-6.
24. *Farra N., Tomeh N., Rozovskaya A., Habash N.* Generalized Character-Level Spelling Error Correction // ACL (2). – 2014. – P. 161-167.
25. *Белозеров А.А., Вахлаков Д.В., Мельников С.Ю., Пересыпкин В.А., Сидоров Е.С.* Технологические аспекты построения системы сбора и предобработки корпусов новостных текстов для создания моделей языка // Известия ЮФУ. Технические науки. – 2016. – № 12 (185). – С. 29-42. DOI: 10.18522/2311-3103-2016-12-2942.

REFERENCES

1. *Birin, D.A., Mel'nikov S.YU., Peresypkin V.A.* Ob effektivnosti sredstv korrektsii iskazhennykh tekstov dlya rezul'tatov raboty sistem raspoznavaniya [About efficiency of means of correction of the distorted texts for results of work of systems of recognition], Superkomp'yuternye tekhnologii (SKT-2018): Materialy 5-y Vserossiyskoy nauchno-tekhnicheskoy konferentsii [Supercomputer technologies (SKT-2018): Materials of the 5th all-Russian scientific and technical conference]: in 2 vol. Vol. 1. Rostov-on-Don; Taganrog: Izd-vo YuFU, 2018, pp. 71-75.
2. *Subramaniam L.V. et al.* A survey of types of text noise and techniques to handle noisy text, *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, July 23-24, 2009, Barcelona, Spain.* DOI: 10.1145/1568296.1568315.
3. *Bassil Y., Alwani M.* Post Editing Error Correction Algorithm for Speech Recognition using Bing Spelling Suggestion, *International Journal of Advanced Computer Science and Applications*, 2012, Vol. 3, No. 2, pp. 95-101.
4. *Feld M., Momtazi S., Freigang F., Klakow D., Müller C*. Mobile texting: can post-ASR correction solve the issues? An experimental study on gain vs. costs, *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, February 14-17, 2012*, pp. 37-40. Lisbon, Portugal. DOI: 10.1145/2166966.2166974.
5. *Evershed J., Fitch K.* Correcting Noisy OCR: Context beats Confusion DATeCH 2014, May 19–20, 2014, Madrid, Spain DOI:10.1145/2595188.2595200.
6. *Lopresti D.P.* Optical character recognition errors and their effects on natural language processing, *International Journal on Document Analysis and Recognition (IJDAR)*, September 2009, Vol. 12, Issue 3, pp. 141–151. DOI: 10.1007/s10032-009-0094-8.
7. *Packer T.L., Lutes J.F., Stewart A.P., Embley D.W., Ringger E.K., Seppi K.D., et al.* Extracting person names from diverse and noisy OCR text, *Proceedings of the fourth workshop on Analytics for noisy unstructured text data AND '10, 2010*, pp. 19-26. DOI 10.1145/1871840.1871845
8. *Kumar A., Lehal G.S.* Automatic Text Correction for Devanagari OCR, *Indian Journal of Science and Technology*, December 2016, Vol. 9 (45). DOI: 10.17485/ijst/2016/v9i45/106372.

9. *Gadde P., Goutam R., Shah R., Bayyarapu H.S., Subramaniam L.V.* Experiments with artificially generated noise for cleansing noisy text, *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, MOCR AND '11,* pp. 4:1-4:8. ACM, 2011.
10. *Dey L., Haque S.K.M.* Studying the effects of noisy text on text mining applications, *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data AND'09*. Barcelona, Spain, 2009, pp. 107-114.
11. *Clark E., Araki K.* Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English, *Procedia - Social and Behavioral Sciences 27, December 2011,* pp. 2-11. DOI: 10.1016/j.sbspro.2011.10.577.
12. *Saloot M.A., Idris N., Mahmud R.* An architecture for Malay Tweet normalization, *Inf. Process. Manag*., 2014, Vol. 50, No. 5, pp. 621-633, DOI: 10.1016/j.ipm.2014.04.009.
13. *Wang A., Kan M.-Y., Andrade D., Onishi T., Ishikawa K.* Chinese Informal Word Normalization: an Experimental Study, *International Joint Conference on Natural Language Processing*, 2013, pp. 127-135. DOI: 10.1007/978-3-319-68612-7_25.
14. *Tursun O., Cakici R.* Noisy Uyghur Text Normalization, Proceedings of the 3rd Workshop on Noisy User-generated Text, Copenhagen, Denmark, September 7, 2017, pp. 85-93. DOI: 10.18653/v1/w17-4412.
15. *Ikeda T., Shindo H., Matsumoto Y.* Japanese Text Normalization with Encoder-Decoder Model, *Proceedings of the 2nd Workshop on Noisy User-generated Text. – Osaka, Japan, December 11, 2016,* pp. 118-126.
16. *Bassil, Y., Alwani, M.* OCR post-processing error correction algorithm using Google's online spelling suggestion, *Journal of Emerging Trends in Computing and Information Sciences*, January 2012, Vol. 3, No. 1.
17. *Спеллер* – Технологии Яндекса. Available at: https://tech.yandex.ru/speller/ (accessed 08 November 2018).
18. *AfterScan* – post-OCR text proofing, advanced spell-checking, automatic correction. Available at: http://www.afterscan.com/ru/ (accessed 08 November 2018).
19. *Turdakov D. i dr.* Texterra: infrastruktura dlya analiza tekstov [Texterra: Infrastructure for text analysis], *Trudy Instituta sistemnogo programmirovaniya RAN* [Proceedings of Institute for system programming of Russian Academy of Sciences], 2014, Vol. 26, Issue 1, pp. 421-438. DOI: 10.15514/ISPRAS-2014-26(1)-18.
20. *Microsoft* Cognitive Services – API Bing проверки орфографии. Available at: https://azure.microsoft.com/ru-ru/services/cognitive-services/spell-check/ (accessed 08 November 2018).
21. *Meshcheryakov R.V.* Struktura sistem sinteza i raspoznavaniya rechi [Structure of speech synthesis and recognition systems], *Izvestiya Tomskogo politekhn. un-ta* [News of Tomsk Polytechnic University], 2009, Vol. 315, No. 5, pp. 127-132.
22. *Smirnov S.V.* Korrektirovka oshibok opticheskogo raspoznavaniya na osnove reytingo-rangovoy modeli teksta [Correction of optical recognition errors based on the rating-rank model of the text], *Trudy SPIIRAN* [SPIIRAS Proceedings], 2014, Issue 4, No. 35, pp. 64-82. DOI: 10.15622/sp.35.5.
23. *Rudakov I.V., Romanov A.S.* Raspoznavanie tekstovogo izobrazheniya s uchetom morfologii slova [Recognition of a text image taking into account the morphology of the word], *Nauka i obrazovanie: nauchnoe izdanie MGTU im. N.E. Baumana* [Science and education: scientific publication of MSTU. N.E. Bauman], 2012, Issue 4, pp. 1-6.
24. *Farra N., Tomeh N., Rozovskaya A., Habash N.* Generalized Character-Level Spelling Error Correction, *ACL (2),* 2014, pp. 161-167.
25. *Belozerov A.A., Vakhlakov D.V., Mel'nikov S.YU., Peresypkin V.A., Sidorov E.S.* Tekhnologicheskie aspekty postroeniya sistemy sbora i predobrabotki korpusov novostnykh tekstov dlya sozdaniya modeley yazyka [Technological aspects of creation of system of gathering and preprocessing of the corpora of news texts to create language models], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2016, No. 12 (185), pp. 29-42. DOI: 10.18522/2311-3103-2016-12-2942.


Статью рекомендовал к опубликованию д.т.н., профессор Р.В. Мещеряков.

**Бирин Дмитрий Анатольевич** – ФГУП НИИ «Квант»; e-mail: melnikov@linfotech.ru; 125438, г. Москва, 4-й Лихачевский пер., 15; зам. директора.

**Пересыпкин Владимир Анатольевич** – ФГУП «НТЦ «Орион»; e-mail: melnikov@linfotech.ru; 127018, г. Москва, ул. Образцова, 38, стр. 1; научный консультант; к.т.н.

**Мельников Сергей Юрьевич** – ООО «Линфо»; e-mail: melnikov@linfotech.ru; 127018, г. Москва, ул. Образцова, 38, стр. 1; тел.: +79037222824; зам. директора; к.ф.-м.н.

**Писарев Илья Александрович** – Южный федеральный университет; e-mail: melnikov@linfotech.ru; 347922, г. Таганрог, ул. Чехова, 2; инженер.

**Цопкало Николай Николаевич** – e-mail: melnikov@linfotech.ru; с.н.с.; к.т.н.

**Birin Dmitrij Anatol'evich** – FGUP NII «Kvant»; e-mail: melnikov@linfotech.ru; 125438, Moscow, 4-j Lihachevskij per., 15; deputy Director.

**Peresypkin Vladimir Anatol'evich** – FGUP "NTC "Orion""; e-mail: melnikov@linfotech.ru; 127018, Moscow, Obrazcova street, 38, str. 1; research consultant; cand. of eng. sc.

**Melnikov Sergey Yur'evich** – OOO "Lingvisticheskie I informatsionye tehnologii" (Limited Liability Company); e-mail: melnikov@linfotech.ru; 127018, Moscow, Obrazcova street, 38, str. 1; deputy Director; cand. of phys.-math. sc.

**Pisarev Il'ya Aleksandrovich** – Southern Federal University; e-mail: melnikov@linfotech.ru; 347922, Taganrog, Chekhova street, 2; engineer.

**Copkalo Nikolaj Nikolaevich** – e-mail: melnikov@linfotech.ru; senior scientist; cand. of eng. sc.

УДК 519.224.22 DOI 10.23683/2311-3103-2018-8-114-135

**А.К. Мельников**

# ПРИМЕНЕНИЕ ТОЧНЫХ И ПРЕДЕЛЬНЫХ ПРИБЛИЖЕНИЙ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ ЗНАЧЕНИЙ СТАТИСТИК ПРИ РЕШЕНИИ ЗАДАЧИ ПО ОБРАБОТКЕ ТЕКСТОВ

*Рассматривается применение предельных и точных приближений распределения вероятностей значений статистик для решения задачи по отбору текстов с определенными статистическими свойствами. Для отбора текстов с равновероятным распределением входящих в них знаков используется статистический критерий согласия, в котором в качестве эталонного распределения тестовой статистики используются его различные приближения. В качестве предельных приближений используются предельные распределения, а в качестве точных приближений - Δ–точные распределения, которые отличающиеся от точных распределений не более чем на заданную величину Δ. Приведены результаты расчета Δ–точных распределений, показаны их отличия от значений предельных распределений для разных статистик. Рассмотрено понятие эффективности обработки по выделению равновероятных текстов, отражающее долю ложно отобранных текстов. Проведено сравнение значений эффективности обработки при применении точных и предельных приближений эталонных распределений тестовых статистик. Показано, что значение эффективности обработки не уменьшается, а во многих случаях растет при применении точного приближения вместо предельного. На основе анализа относительной эффективностью критериев и методов исследования их асимптотического поведение при различных ограничениях, для сравнение статистических критериев, использующих одинаковую тестовую статистику но разные её эталонные распределения вводится понятие относительной эффективности распределения, показывающее во сколько раз увеличится количество ложно отобранных текстов при применении в качестве эталонного распределения критерия того или иного распределения. Показана функциональная связь между понятиями эффективность обработки и относительная эффективность распределений. В условиях доступности высокопроизводительных вычислительных средств, позволяющих проводить расчеты Δ–точных распреде-*