Раздел III. Обработка информации в распределенных, реконфигурируемых и нейросетевых системах

УДК 004.931 DOI 10.18522/2311-3103-2021-7-130-142

Д.В. Вахлаков, А.В. Германович, С.Ю. Мельников, В.А. Пересыпкин, Н.Н. Цопкало

О ТОЧНОСТИ И ТРУДОЕМКОСТИ МНОГОЭТАПНОГО МЕТОДА КОРРЕКЦИИ ИСКАЖЕННЫХ ТЕКСТОВ В ЗАВИСИМОСТИ ОТ СТЕПЕНИ ИСКАЖЕНИЯ

Одним из основных факторов, существенно затрудняющих понимание, перевод и анализ текстов, полученных при автоматическом распознавании речи или изображений текстов, являются содержащиеся в них искажения в виде ошибочных символов, слов и словосочетаний. До недавнего времени не существовало эффективных программных средств коррекции текстов со значительными искажениями, хотя эта задача является актуальной как для русского, так и для других распространенных языков в условиях активного использования систем распознавания в перспективных системах дополненной реальности. Авторами был предложен новый многоэтапный метод коррекции искаженных текстов, значимо повышающий точность коррекции (количество правильно скорректированных слов в тексте) и основанный на последовательном определении ошибок и их исправлении. В настоящей работе оцениваются точность и трудоемкость предложенного метода коррекции искаженных текстов при различных уровнях искажений, определяется его место среди других современных подходов к коррекции. Наиболее характерными ошибками систем распознавания являются: – замена слова на похожее по звучанию или графическому написанию; – замена нескольких слов на одно; – замена одного слова несколькими; – пропуск слов; – вставка или удаление коротких слов (в т.ч. предлогов и союзов). В результате распознавания получается текст, имеющий искажения и состоящий, в основном, из словарных слов, в том числе и в местах искажений. При большом количестве искажений тексты становятся практически нечитаемыми. В связи с тем, что подобрать в необходимом количестве тексты с широким диапазоном уровней искажений по результатам реального машинного распознавания речи и изображений текстов представляется проблематичным, использовалось программное моделирование искажений. Предложена и программно реализована методика искажений текста, моделирующая результаты работы систем распознавания в широком диапазоне искажений, в необходимом количестве подготовлены искаженные тексты. При работе предложенного многоэтапного метода коррекции искаженными считаются несловарные словоформы и словоформы, вероятность появления которых в тексте в соответствии с выбранной вероятностной моделью текста меньше заданного порога. Для них строится список возможных вариантов слов, в который попадают только те словоформы из словаря, которые находятся от исследуемого слова на определенном расстоянии Левенштейна. Скорректированный текст из вариантов слов получается в результате поиска наиболее вероятной цепочки словоформ. Метод коррекции состоит из нескольких этапов, на каждом этапе корректируются лишь те фрагменты текста, которые остались искаженными после предыдущего этапа коррекции. По результатам проведенных экспериментов по коррекции искаженных текстов сделан вывод, что предложенный метод коррекции показал хорошие результаты со средним значением F_1*-меры >50 % в диапазоне искажений от 0 до 75 %. Эксперты-лингвисты подтвердили плодотворность предложенного подхода к коррекции и его предпочтительность по сравнению с другими современными подходами, зафиксировав, что при количестве искажений*

<50 % скорректированный текст читается с гораздо меньшими усилиями, чем искаженный, а при количестве искажений до 70% слов скорректированный текст еще позволяет выделить полезную информацию о содержании текста.

Многоэтапный метод коррекции искаженных текстов; модель языка; расстояние Левенштейна; полнота и точность коррекции; F1-мера; WER; CER; эксперты-лингвисты.

D.V. Vakhlakov. A.V. Germanovich, S.Yu. Melnikov, V.A. Peresypkin, N.N. Copkalo

ON THE ACCURACY AND COMPLEXITY OF THE MULTI-STAGE METHOD FOR CORRECTING DISTORTED TEXTS DEPENDING ON THE DEGREE OF DISTORTION

One of the main factors that significantly complicate the understanding, translation and analysis of texts obtained by automatic recognition of speech or images of texts is the presence of distortions in the form of erroneous symbols, words and phrases. Until recently, there were no effective software tools for correcting texts with significant distortions, although this task is relevant both for Russian and other common languages in the context of the active use of recognition systems in advanced augmented reality systems. The authors proposed a new multi-stage method for correcting distorted texts, which significantly increases the accuracy of the correction (in terms of the number of correctly corrected words in the text) and is based on the sequential detection of errors and their correction. In this paper, we evaluate the accuracy and computational complexity of the proposed method for correcting distorted texts at various levels of distortion, and determine its place among other modern approaches to correction. The most typical errors of recognition systems are: – replacing a word with a similar sound or graphic spelling; – replacing several words with one; – replacing one word with several; – omission of words; – insertion or deletion of short words (including prepositions and conjunctions). As a result of recognition, a distorted text is obtained, which consists mainly of dictionary words, even in places of distortion. With a large number of distortions, the texts become almost unreadable. Due to the fact that it is problematic to select texts with a wide range of distortion levels in the required amount based on the results of real machine recognition of speech and images of texts, software modeling of distortions was used. A text distortion technique has been proposed and implemented that simulates the results of recognition systems in a wide range of distortions; distorted texts have been prepared in the required amount. Within the framework of the proposed multi-stage correction method, non-dictionary word forms and words are considered distorted if the probability of their occurrence in the text in accordance with the chosen language model is less than a given threshold. For such distorted words, a list of possible variants of words is built, which includes only those word forms from the dictionary that are at a certain Levenshtein distance from the word under study. The corrected text from the tables of word variants is obtained by searching for the most probable chain of word forms. The correction method consists of several stages, at each stage only those fragments of the text that remain distorted after the previous stage are corrected. According to the results of the experiments on the correction of distorted texts, it was concluded that the proposed correction method showed good results with an average value of F-measure >50 % in the distortion range from 0 to 75 %. Linguistic experts confirmed the fruitfulness of the proposed approach to correction and its preference over other modern approaches, fixing that with a level of distortion of up to 50 % of words, the corrected text is read with much less effort than a distorted one, and with a level of distortion of up to 70% of words, the corrected text also allows you to highlight useful information about the content.

Multi-stage method of text correction; language model; Levenshtein distance; completeness and accuracy of correction; F1-measure; WER; CER; linguistic experts.

Введение. Основным фактором, существенно затрудняющим понимание и перевод текстов, полученных при машинном распознавании речи или изображений текстов, являются содержащиеся в них искажения в виде ошибочных символов, слов и даже словосочетаний [1–3]. При значительном количестве искажений автоматическая обработка таких текстов практически невозможна.

До недавнего времени не существовало эффективных программных средств коррекции текстов со значительными искажениями (или сильно искаженных текстов), хотя эта задача является актуальной как для русского, так и для других распространенных языков в условиях активного использования систем распознавания в перспективных системах дополненной реальности [4, 5]. В работе [6] предложен новый многоэтапный метод коррекции искаженных текстов (в том числе, и сильно искаженных), значимо повышающий точность коррекции и основанный на последовательном определении ошибок и их исправлении. В настоящей работе оцениваются точность и вычислительные затраты предложенного метода коррекции искаженных текстов при различных уровнях искажений, определяется его место среди других современных подходов к коррекции.

1. Современные подходы к коррекции искажений в текстах. Наличие искажений в текстах значительно снижает эффективность их автоматической обработки. В обзоре [7] приведен перечень практических приложений, в которых возникают искаженные тексты. К таким приложениям, в частности, относятся: коррекция текстов, обработка коротких сообщений в социальных сетях, оптическое распознавание символов, распознавание речи, распознавание рукописного текста, машинный перевод, добывание информации из текста, обработка текстовых запросов, классификация и аннотирование текстов, а также задачи анализа оперативной обстановки с помощью систем дополненной реальности, использующих перевод речи и изображений документов с иностранных языков [8], [9].

В [10] показано, что орфографические ошибки в текстах патентов приводят к значительному ухудшению точности обработки поисковых запросов, и предложена специальная процедура обработки текстов патентов для повышения эффективности патентного поиска.

Специфические методы пост-обработки текстов, полученных в результате оптического распознавания, анализируются в [11]. Способы борьбы с искажениями текстов, которые возникают при наборе на клавиатуре, излагаются в [12] и [13].

Для моделирования искажений текстов предлагаются различные модели случайности. Так, в [14] рассматриваются следующие типы случайных искажений слов в тексте: бернуллиевский шум, гауссов шум, а также введенный авторами «состязательный» шум. Для повышения устойчивости в задаче классификации предложений предлагается обучение на текстах, подвергнутых рассмотренными искажениями рассмотренных типов. В [15] предложен способ моделирования искажений текстов, позволяющий получать искаженные тексты, близкие к тем, которые являются результатом работы систем распознавания. Способ использует взвешенную смесь случайных замен символов и случайных замен слов на близкие по расстоянию Левенштейна.

В [6] предложен многоэтапный метод коррекции искаженных текстов (в том числе, и сильно искаженных), основанный на последовательном определении ошибок и их исправлении. Отметим также возможность привлечения экспертов-лингвистов для объективизации и уточнения оценок качества работы автоматических систем обработки искаженных текстов [3].

2. Устойчивость языковых моделей к искажениям. В [16] для повышения устойчивости нейросетевых моделей к шумам предложены модификации схем вложения слов (words embedding) для следующих задач: классификация текстов, распознавание поименованных сущностей, извлечение аспектов. В [17] задача коррекции результата оптического распознавания рассматривается как задача перевода с одного языка на другой, и для нее применяется модель трансформера на уровне предложений. Достигнуто посимвольное улучшение точности распознавания на 29.4 %.

Заметное влияние на исследования в области обработки текстов оказала работа [18], в которой показано, что символьные нейросетевые модели, которые используются для машинного перевода, как правило, не способны справляться ни с естественно возникающими искажениями в тексте, ни с искусственно внесенными. Рассмотрены четыре типа случайных искажений, связанных с перестановками букв внутри слова (1 – транспозиция соседних букв, 2 – случайное нарушение порядка следования букв в слове, за исключением первой и последней, 3 – случайная перестановка букв в пределах слова и 4 – случайная замена буквы на другую). Исследования проводились с текстами на французском, немецком и чешском языках, для оценки качества перевода использован показатель BLEU. Показано, что системе автоматического перевода Google Translate не удается переводить даже умеренно искаженные тексты, легко понимаемые человеком.

В [19] анализируется, как на точность машинного перевода влияет зашумленность обучающих данных. Рассмотрено несколько вариантов зашумления обучающих параллельных корпусов текстов. Показано, что нейросетевые системы автоматического перевода (NMT, Neural Machine Translation) менее устойчивы к зашумлению обучающих данных, чем системы автоматического перевода, построенные на статистических принципах (SMT, Statistical Machine Translation).

Предварительно обученные нейросетевые языковые модели, такие как BERT (Bidirectional Encoder Representations from Transformers, [20]) в настоящее время обеспечивают наивысшее качество решения многих задач в области вычислительной лингвистики, таких как аннотирование, распознавание поименованных сущностей, машинный перевод и др. В [21] рассматривается эффективность BERT в задачах анализа искаженных текстов. В качестве искажений выступают случайные опечатки, т.е. замена символа на другой символ, расположенный рядом на клавиатуре типа QWERTY. Рассматривался диапазон искажений от 0 до 22.5%. Показано, что с ростом уровня ошибок эффективность BERT резко падает. В частности, в задачах оценки тональности и определения близости предложений при уровне символьных ошибок в 15–17 % результат сопоставим со случайным выбором возможных ответов. Авторы выдвигают предположение, что точность работы BERT может повыситься, если входные тексты перед точной настройкой BERT предварительно обработать системой коррекции ошибок. Другим вариантом решения этой проблемы может стать изменение архитектуры BERT для повышения его устойчивости к шуму.

В работах [22–24] отмечается, что современные нейросетевые методы машинного перевода неустойчивы к зашумлению текстов. Предложены подходы к искусственному зашумлению корректных текстов, которые бы обеспечивали их похожесть на некорректные тексты из социальных сетей и при использовании в качестве обучающих корпусов позволяли бы улучшить качество машинного перевода таких текстов.

В [25, 26] и ряде других предложены различные способы повышения надежности перевода зашумленных текстов.

В целом, можно сказать, что прикладные системы, использующие языковые модели, весьма чувствительны к шуму во входных текстах, например, к ошибкам в написании слов. Разрабатываемые сегодня подходы для описания устойчивости таких систем к зашумлению входных данных используют, во-первых, только самые простые типы искажений, во-вторых, ограничиваются невысокими уровнями искажений. Кроме того, недостаточно внимания уделяется росту вычислительных затрат, необходимых таким системам для работы в условиях искажений.

В настоящей работе проведены экспериментальные оценки точности и трудоемкости (скорость работы на современных вычислительных средствах) предложенного метода коррекции искаженных текстов в максимально широком диапазоне искажений.

3. Методика искажений текстов. В связи с тем, что подобрать в необходимом количестве тексты с широким диапазоном уровней искажений по результатам реального машинного распознавания речи и изображений текстов представляется проблематичным, использовалось программное моделирование искажений. По результатам проведенного анализа ошибок систем распознавания систематизированы наиболее характерные типы ошибок, а именно: – замена слова на похожее по звучанию или графическому написанию; – замена нескольких слов на одно; – замена одного слова несколькими; – пропуск слов; – вставка или удаление коротких слов (частиц, предлогов и союзов). При распознавании изображений текста возникают ошибки, при которых часть символов заменяется на близкие по написанию. В связи с использованием для восстановления текста лингвистических словарей в результате получается текст, имеющий искажения и состоящий, в основном, из словарных слов, в том числе и в местах искажений.

Приведем краткое описание двухэтапного алгоритма, моделирующего целевые искажения систем машинного распознавания. 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 без исключения (в отличие от [15], где исключались символы, составляющие 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.

4. Метод коррекции искаженных текстов. В соответствии с многоэтапным методом ([6]) коррекции искаженных текстов искаженными считаются несловарные словоформы и словоформы, вероятность появления которых в тексте в соответствии с выбранной вероятностной моделью меньше заданного порога ([27]). После установки признака искаженности для отдельных слов происходит распространение этого признака на их сочетания, т.е. выделяются искаженные фрагменты текста. Для них строится список возможных вариантов слов, в который попадают только те словоформы из словаря, которые находятся от исследуемого слова на определенном расстоянии Левенштейна (1,2,…,*k*), которое характеризуют минимальное количество изменений (вставок, замен и удалений) слов, необходимых для преобразования одной последовательности слов в другую. Скорректированный текст получается в результате обработки колонок слов с помощью алгоритма динамического программирования с заданной глубиной зависимости между словами и построения наиболее вероятной цепочки слов на основе выбранной N-граммной вероятностно-статистической модели на словах.

Метод коррекции состоит из *k* этапов, на каждом этапе корректируются лишь те фрагменты текста, которые остались искаженными после предыдущего этапа коррекции. Количество этапов является параметром алгоритма, который определяет максимальное расстояние Левенштейна и задается в зависимости от степени искаженности текста.

Эффективность коррекции оценивается с позиций точности и полноты (F_1 мера), и скорости (количество скорректированных слов в секунду). F_1 мера это гармоническое среднее точности A и полноты R коррекций искажённого текста с одинаковым весом:

$$F_1 = \frac{2AR}{(A+R)}.$$

Полнота коррекции R рассчитывается как отношение количества верно скорректированных слов $W(T)$ к количеству слов в искажённых фрагментах $W(E_1)$:

$$R = \frac{W(T)}{W(E_1)},$$

а точность коррекции A – как обратное отношение количества слов в неверных коррекциях $W(F)$ к количеству слов в искажённых фрагментах $W(E_1)$:

$$A = 1 - \frac{W(F)}{W(E_1)}.$$

5. Описание проведенных экспериментов. Из современного новостного корпуса ([28]) текстов английского языка были отобраны 300 текстов объемом около 1500 символов (приблизительно одна страница) каждый. По представленной в п. 3 методике моделирования искажений эти тексты искажались с разным уровнем искажений (от 1 до 99 % искаженных слов). Для удобства были введены 48 градаций уровней искажений, с более крупным шагом в середине исследуемого диапазона. Для каждой градации были независимо друг от друга искажены отобранные 300 текстов и для каждого полученного искаженного текста подсчитаны значения *WER* (Word Error Rate) и *CER* (Character Error Rate), определяемые следующим образом.

Пусть в тексте объемом N слов (n символов) проведены I (i) вставок, D (d) удалений, S (s) замен слов (символов). Тогда

$$WER = \frac{I + D + S}{N} \bullet 100\%,$$

$$CER = \frac{i + d + s}{n} \bullet 100\%.$$

Для каждой градации были подсчитаны средние по всем текстам частоты ошибочных слов *WER* и ошибочных символов *CER*. Результаты расчетов сведены в табл. 1 и представлены на рис. 1.

Таблица 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
WER	1.11	2.17	3.22	4.26	5.36	6.28	7.19	8.31	9.33	10.21	11.04	12.	13.12	14.04	14.92	19.75
CER	0.18	0.35	0.52	0.69	0.88	1.03	1.19	1.37	1.54	1.7	1.85	2.02	2.21	2.39	2.54	3.43
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
WER	24.78	29.73	34.97	39.87	45.1	50.06	54.87	60.09	65.03	69.9	75.09	80.01	84.89	89.76	93.82	95.14
CER	4.41	5.45	6.57	7.71	9.	10.34	11.71	13.31	15.08	16.89	19.1	21.6	24.65	28.3	32.61	36.2
	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
WER	95.47	95.89	96.23	96.64	96.99	97.22	97.52	97.82	98.05	98.26	98.46	98.61	98.74	98.9	98.97	99.06
CER	37.69	39.34	41.06	43.03	45.	47.15	49.47	51.87	54.6	57.25	60.21	63.21	66.27	69.51	72.67	75.91

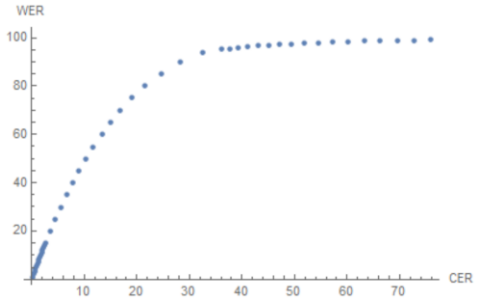

Рис. 1. Значения CER и WER и для 48 градаций искажений

Как видно из рис. 1, вначале (до *CER*<10 и *WER*<50) наблюдается практически линейный характер зависимости между *CER и WER* и высокая скорость роста значений *WER*, затем скорость роста значений снижается и с *CER*>35 значение *WER* изменяется минимально.

Описанный в п. 4 метод коррекции искаженных текстов программно реализован на вычислителе следующей конфигурации: процессор – Intel (R) Xeon (R) CPU E5-2699 v4 @ 2.20GHz, 44 ядра, ОЗУ - 250 ГБ, с ОС Windows Server 2012 R2 Standard, x6. В качестве модели английского языка была использована вероятностно-статистическая 4-граммная модель Маркова на словах со сглаживанием Кнесера-Нея. Модель была построена на корпусе англоязычных общественно-политических текстов объемом 200 млн. слов.

Максимальное значение расстояния Левенштейна составило 4, т.е. алгоритм коррекции состоял из 4-х этапов.

6. Анализ результатов коррекции. В табл. 2 в качестве примера приведен фрагмент одного из 300 использованных текстов в неискаженном варианте и в пяти вариантах искажений с последующей коррекцией.

Таблица 2

Неискаженный текст		A man was killed and at least seven other people were wounded in separate shootings Monday on the city's South, North and West sides, said officials.
WER =10.21	искаженный	A man was killed and at least oven other peole were woznded in separate shootings Mnday on the city's South, North aad West sids, said officialsa.
	скорректированный	A man was killed and at least oven other people were wounded in separate shootings Monday on the city's South, North and West sids, said officials.
WER =29.73	искаженный	A man was tilled abnd at laste sven mthr people were wounjded in separate hoovtings Monday onte riqty's South, North and Wes sides, said officials..
	скорректированный	A man was killed and at last sven mohr people were wounded in separate shootings Monday one city's South, North and Wes sides, said officials.
WER =50.06	искаженный	A may was kiled andj at yeast seden other peapleywere aounded in sesarate shwotings Monday on the ciy's Soiuth, North awd West sizes,msaif officgals.
	скорректированный	A may was killed and at yeast seven other police wounded in separate shootings Monday on the city's South, North awd West sizes,saif officials.

WER =69.9	искаженный	A man wls villedanyat yeast xever oter peope wdr rounded ins separated shootiunga Moday octhni rity' South, Northoand West sidwes, sqmid offfcals.
	скорректированный	A man wls dried yeast ever other people dr rounded ins separated shooting Today ohtani city' South, Northland West sides, squid officers ap.
WER =89.76	искаженный	A mjez gas rillesd apdnait lest semun roqhesr people werye buunred hin seprtexhotidngaMoedaay oy the city'sca Soulh, Notrthsand Westzides, sazd offcials.
	скорректированный	A new gas filled apart lest segun after people were burned hin Seprtexhotidngamoedaay on the city'sca South, Northland Westside, said officials.

Приведенные в табл. 2 результаты демонстрируют выводы экспертов-лингвистов, что при *WER* < 50 скорректированный текст читается с гораздо меньшими усилиями, чем искаженный. При 50 < *WER* < 70 скорректированный текст еще позволяет выделить некоторую полезную информацию о содержании текста.

При *WER* > 70 верно скорректированными оказываются отдельные слова, которых недостаточно для передачи смысла всего текста.

Эксперты-лингвисты отмечают, что ошибки алгоритма коррекции часто связаны с наличием в искаженном тексте словарных искажений, то есть таких, которые слово при искажении переводят в другое слово из словаря. Приведем несколько примеров.

Пример скорректированного текста из градации ***WER=50.06***: *A may was killed*. Если допустить, что *may* здесь означает «май» (хотя название месяцев в английском языке пишутся с прописной буквы), то тогда откорректированная фраза должна означать «Май был убит» (вариант перевода may глаголом «мочь» здесь не подходит). Проведенная коррекция не поменяла *may* на *man* («мужчина»), хотя понятно, что убит был человек; предположить здесь некое переносное значение (май «был убит» в смысле «прошел зря», «не задался») не позволяет контекст: далее упоминаются «еще семеро, которые были ранены».

Другим характерным типом ошибок алгоритма коррекции являются ошибки при разделении слитно написанных искаженных слов. Так, последовательность букв *onte* из градации ***WER=29.73*** не заменена двумя словами (*on+the*), как, возможно, поступил бы человек, а просто исправлена на слово *one* из словаря (ограничившись одношаговой трансформацией по Левенштейну вместо двушаговой). Показателен в этом отношении также пример последовательности букв *peaplewere* из градации ***WER=50.06***, которое ошибочно было исправлено на *police*, вместо разделения ее на два слова – *people+were*, что сделал бы человек, учтя узкий контекст (предшествующее слово *other* и последующее причастие прошедшего времени *wounded* вместе со вспомогательным глаголом *were*). В примерах из ***WER=89.76*** и ***WER=69.9*** *Northoand* и *Notrthsand* были ошибочно заменены на словарное слово *Northland* («северные земли»), вместо того чтобы разделить эти «слова» на *North* + *and*, как скорее всего поступил бы человек.

Эксперты также отмечают, что, алгоритмы коррекции, применяемые для одного языка или группы языков, могут не подходить либо не полностью подходить для других языков в силу их особенностей, что требует выработки специфических методов для каждого конкретного языка.

На графиках на рис. 2 показаны зависимости достигаемой точности коррекции в терминах F_1-меры от величины исходных искажений. Максимальная точность коррекции соответствует величине 30 % по *WER* или 6 % по *CER*. Величины F_1-меры, большие 50%, соответствуют диапазону 0–75 % по *WER* или 0–20 % по *CER*.

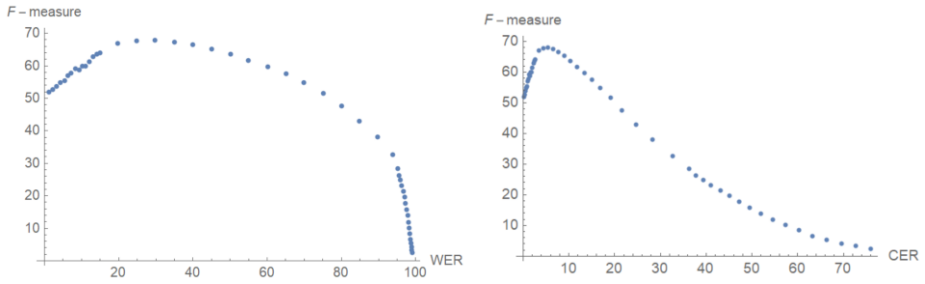

Рис. 2. Зависимость F_1-меры от WER (слева) и CER (справа)

На графиках на рис. 3 показана средняя скорость коррекции в зависимости от величины искажения. Отметим, что если при *WER*>20 (*CER*>4) скорость коррекции является стабильно низкой, то при меньших уровнях искажений средняя скорость нестабильна. По всей видимости, при таких уровнях искажений на скорость коррекции существенно влияет конкретное содержание текста и характер искаженных слов.

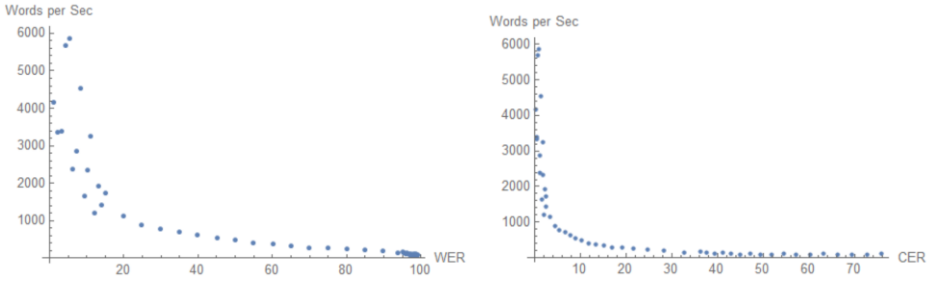

Рис. 3. Зависимости достигнутой скорости коррекции, в словах в секунду, от WER (слева) и CER (справа)

Выводы. Для определения плодотворности предложенного метода коррекции искаженных текстов и сравнения с другими современными подходами к коррекции, экспериментально оценена точность и трудоемкость его работы в широком диапазоне значений искажений. Полученные результаты показывают работоспособность этого метода коррекции в различных областях применения, в том числе в системах дополненной реальности.

Метод коррекции основан на последовательном многоэтапном исправлении искаженных слов с использованием 4-х граммной вероятностно-статистической модели Маркова.

Предложена и программно реализована методика искажений текста, моделирующая результаты работы систем распознавания речи и изображений текстов в широком диапазоне искажений от 1 до 99 % по *WER*, в необходимом для проведения экспериментов количестве подготовлены искаженные тексты.

По результатам проведенных экспериментов по коррекции искаженных текстов сделан вывод, что предложенный метод коррекции показал хорошие результаты со средним значением F_1-меры >50 % в диапазоне искажений от 0 до 75 % по *WER*.

Эксперты-лингвисты подтвердили плодотворность предложенного подхода к коррекции и его предпочтительность по сравнению с другими современными подходами, зафиксировав, что при *WER*<50 % скорректированный текст читается с гораздо меньшими усилиями, чем искаженный, а при 50 %<*WER*<70 % скорректированный текст еще позволяет выделить полезную информацию о содержании текста.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Мещеряков Р.В.* Структура систем синтеза и распознавания речи // Известия Томского политехн. ун-та. – 2009. – Т. 315, № 5. – С. 127-132.
2. *Смирнов С.В.* Корректировка ошибок оптического распознавания на основе рейтингоранговой модели текста // Тр. СПИИРАН. – 2014. – Вып. 4. – № 35. – С. 64-82.
3. *Германович А.В.,Мельников С.Ю., Пересыпкин В.А., Сидоров Е.С., Цопкало Н.Н.* Информационные измерения языка. Программная система оценки читаемости искаженных текстов // Известия ЮФУ. Технические науки. – 2019. – № 8. – С. 6-18.
4. www.topwar.ru > 18316 – pehotnaja-sistema-dopolnennoj-realnosti-IVAS (США). 29.03.2021.
5. www.tadviser.ru > index php / Статья Компьютерное_зрение_технологии_рынок_перспективы. 26.06.2019.
6. *Вахлаков Д.В., Мельников С.Ю., Пересыпкин В.А.* Многоэтапный метод автоматической коррекции искаженных текстов // Известия ЮФУ. Технические науки. – 2020. – № 7. – С. 35-45.
7. *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.
8. https://www.ldc.upenn.edu/collaborations/current-projects/madcat.
9. *Strassel S., Friedman L., Ismael S., Brandschain L.* New Resources for Document Classification, Analysis and Translation Technologies // Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008.
10. *Stein B., Hoppe D., Gollub T.* The impact of spelling errors on patent search // In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012). – P. 570-579.
11. *Nguyen T., Jatowt A., Coustaty M., Doucet A.* Survey of Post-OCR Processing Approaches // ACM Comput. Surv. 54, 6, Article 124 (July 2021). – 37 p.
12. *Ghosh S., Kristensson P.* Neural Networks for Text Correction and Completion in Keyboard Decoding // arXiv:1709.06429, 2017.
13. *Рыбанов А.А., Филиппова Е.М., Свиридова О.В., Федотова Л.А.* Система количественных показателей мониторинга за процессом развития навыка ввода информации // Педагогическая информатика. – 2020. – № 1. – С. 136-142.
14. *Zhang D., Yang Z.* Word Embedding Perturbation for Sentence Classification // CoRR preprint arXiv:1804.08166, 2018.
15. *Бирин Д.А., Мельников С.Ю., Пересыпкин В.А., Писарев И.А., Цопкало Н.Н.* Об эффективности средств коррекции искаженных текстов в зависимости от характера искажений // Известия ЮФУ. Технические науки. – 2018. – № 8. – С. 104-114.
16. *Malykh V.* Robust-to-Noise Models in Natural Language Processing Tasks // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, Italy, July 28 - August 2, 2019. – P. 10-16.
17. *Soper E., Fujimoto S., Yu Y.* BART for Post-Correction of OCR Newspaper Text // Proceedings of the 2021 EMNLP Workshop W-NUT: The 7th Workshop on Noisy User-generated Text. November 11, 2021. – P. 284-290.
18. *Belinkov Y., Bisk Y.* Synthetic and natural noise both break neural machine translation // arXiv:1711.02173, 2017.
19. *Khayrallah H., Koehn P.* On the Impact of Various Types of Noise on Neural Machine Translation // In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. – 2018. – P. 74-83.
20. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv:1810.04805, 2018.
21. *Kumar A., Makhija P., Gupta A.* Noisy Text Data: Achilles' Heel of BERT // Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text. – P. 16-21.
22. *Vaibhav, Singh S., Stewart C., Neubig G.* Improving Robustness of Machine Translation with Synthetic Noise // arXiv:1902.09508, 2019.
23. *Niu X., Mathur P., Dinu G., Al-Onaizan Y.* Evaluating Robustness to Input Perturbations for Neural Machine Translation // arXiv:2005.00580, 2020.

24. *Karpukhin V., Levy O., Eisenstein J., Ghazvininejad M.* Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation // arXiv:1902.01509, 2019.
25. *Li Z., Rei M., Specia L.* Visual Cues and Error Correction for Translation Robustness // arXiv:2103.07352, 2021.
26. *Riabi A., Sagot B., Seddah D.* Can Character-based Language Models Improve Downstream Task Performance in Low-Resource and Noisy Language Scenarios? // Proceedings of the 2021 EMNLP Workshop W-NUT: The 7th Workshop on Noisy User-generated Text. November 11, 2021. – P. 423-436.
27. *Мельников С.Ю., Пересыпкин В.А.* О применении вероятностных моделей языка для обнаружения ошибок в искаженных текстах // Вестник компьютерных и информационных технологий. – 2016. – № 5. – С. 29-34.
28. *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. – С. 29-42.

REFERENCES

1. *Meshcheryakov R.V.* Struktura sistem sinteza i raspoznavaniya rechi [The structure of speech synthesis and recognition systems], *Izvestiya Tomskogo politekhn. un-ta* [Izvestiya Tomsk Polytechnic University], 2009, Vol. 315, No. 5, pp. 127-132.
2. *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], *Tr. SPIIRAN* [Proceedings of SPIIRAN], 2014, Issue 4, No. 35, pp. 64-82.
3. *Germanovich A.V.,Mel'nikov S.Yu., Peresypkin V.A., Sidorov E.S., Tsopkalo N.N.* Informatsionnye izmereniya yazyka. Programmnaya sistema otsenki chitaemosti iskazhennykh tekstov [Information dimensions of language. Software system for assessing the readability of distorted texts], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2019, No. 8, pp. 6-18.
4. www.topwar.ru > 18316 – pehotnaja-sistema-dopolnennoj-realnosti-IVAS (SShA) [www.topwar.ru > 18316 – pehotnaja-sistema-dopolnennoj-realnost-IVAS (USA)]. 29.03.2021.
5. www.tadviser.ru > index php / Stat'ya Komp'yuternoe_zrenie_tekhnologii_rynok_perspektivy [www.tadviser.ru > index php / Article Computer_view_technology_market prospects]. 26.06.2019.
6. *Vakhlakov D.V., Mel'nikov S.Yu., Peresypkin V.A.* Mnogoetapnyy metod avtomaticheskoy korrektsii iskazhennykh tekstov [Multi-stage method of automatic correction of distorted texts], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2020, No. 7, pp. 35-45.
7. *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.
8. Available at: https://www.ldc.upenn.edu/collaborations/current-projects/madcat.
9. *Strassel S., Friedman L., Ismael S., Brandschain L.* New Resources for Document Classification, Analysis and Translation Technologies, *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*.
10. *Stein B., Hoppe D., Gollub T.* The impact of spelling errors on patent search, *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pp. 570-579.
11. *Nguyen T., Jatowt A., Coustaty M., Doucet A.* Survey of Post-OCR Processing Approaches, *ACM Comput. Surv*. 54, 6, Article 124 (July 2021), 37 p.
12. *Ghosh S., Kristensson P.* Neural Networks for Text Correction and Completion in Keyboard Decoding, *arXiv:1709.06429, 2017*.
13. *Rybanov A.A., Filippova E.M., Sviridova O.V., Fedotova L.A.* Sistema kolichestvennykh pokazateley monitoringa za protsessom razvitiya navyka vvoda informatsii [A system of quantitative indicators for monitoring the process of developing the information input skill], *Pedagogicheskaya informatika* [Pedagogical informatics], 2020, No. 1, pp. 136-142.
14. *Zhang D., Yang Z.* Word Embedding Perturbation for Sentence Classification, *CoRR preprint arXiv:1804.08166, 2018*.

15. *Birin D.A., Mel'nikov S.Yu., Peresypkin V.A., Pisarev I.A., TSopkalo N.N.* Ob effektivnosti sredstv korrektsii iskazhennykh tekstov v zavisimosti ot kharaktera iskazheniy [On the effectiveness of the means of correction of distorted texts depending on the nature of the distortion], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2018, No. 8, pp. 104-114.
16. *Malykh V.* Robust-to-Noise Models in Natural Language Processing Tasks, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop.* Florence, Italy, July 28 - August 2, 2019, pp. 10-16.
17. *Soper E., Fujimoto S., Yu Y.* BART for Post-Correction of OCR Newspaper Text, *Proceedings of the 2021 EMNLP Workshop W-NUT: The 7th Workshop on Noisy User-generated Text. November 11, 2021,* pp. 284-290.
18. *Belinkov Y., Bisk Y.* Synthetic and natural noise both break neural machine translation, *arXiv:1711.02173, 2017*.
19. *Khayrallah H., Koehn P.* On the Impact of Various Types of Noise on Neural Machine Translation, *In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation,* 2018, pp. 74-83.
20. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding, *arXiv:1810.04805, 2018*.
21. *Kumar A., Makhija P., Gupta A.* Noisy Text Data: Achilles' Heel of BERT, *Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text,* pp. 16-21.
22. *Vaibhav, Singh S., Stewart C., Neubig G.* Improving Robustness of Machine Translation with Synthetic Noise, *arXiv:1902.09508, 2019*.
23. *Niu X., Mathur P., Dinu G., Al-Onaizan Y.* Evaluating Robustness to Input Perturbations for Neural Machine Translation, *arXiv:2005.00580, 2020*.
24. *Karpukhin V., Levy O., Eisenstein J., Ghazvininejad M.* Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation, *arXiv:1902.01509, 2019.*
25. *Li Z., Rei M., Specia L.* Visual Cues and Error Correction for Translation Robustness, *arXiv:2103.07352, 2021*.
26. *Riabi A., Sagot B., Seddah D.* Can Character-based Language Models Improve Downstream Task Performance in Low-Resource and Noisy Language Scenarios?, *Proceedings of the 2021 EMNLP Workshop W-NUT: The 7th Workshop on Noisy User-generated Text. November 11, 2021,* pp. 423-436.
27. *Mel'nikov S.Yu., Peresypkin V.A.* O primenenii veroyatnostnykh modeley yazyka dlya obnaruzheniya oshibok v iskazhennykh tekstakh [On the application of probabilistic language models to detect errors in distorted texts], *Vestnik komp'yuternykh i informatsionnykh tekhnologiy* [Bulletin of Computer and Information Technologies], 2016, No. 5, pp. 29-34.
28. *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 building a system for collecting and preprocessing news text corpora to create language models], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2016, No. 12, pp. 29-42.

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

Вахлаков Дмитрий Викторович – ФГУП «НТЦ «Орион»; e-mail: melnikov@linfotech.ru; г. Москва, Россия; научный сотрудник.

Пересыпкин Владимир Анатольевич – e-mail: melnikov@linfotech.ru; научный консультант; д.т.н.

Германович Андрей Валерьевич – Московский государственный университет им. М.В. Ломоносова, Институт стран Азии и Африки; e-mail: melnikov@linfotech.ru; г. Москва, Россия; доцент кафедры арабской филологии; к.и.н.

Мельников Сергей Юрьевич – ООО «Линфо»; e-mail: melnikov@linfotech.ru; г. Москва, Россия; тел.: +79037222824; зам. директора; к.ф.-м.н.

Цопкало Николай Николаевич – Южный федеральный университет; e-mail: melnikov@linfotech.ru; г. Таганрог, Россия; с.н.с.; к.т.н.

Vakhlakov Dmitriy Viktorovich – FGUP “NTC “Orion””; e-mail: melnikov@linfotech.ru; Moskow, Russia; researcher.

Peresypkin Vladimir Anatol’evich – e-mail: melnikov@linfotech.ru; research consultant; dr.. of eng. sc.

Germanovich Andrey Valer’evich – Moscow State University, Institute of Asian and African Studies; e-mail: melnikov@linfotech.ru; Moskow, Russia; assistant professor at the arabic philology department; cand. of history. sc.

Melnikov Sergey Yur’evich – OOO “Lingvisticheskie I informatsionye tehnologii” (Limited Liability Company); e-mail: melnikov@linfotech.ru; Moskow, Russia; deputy director; cand. of phys. and math. sc.

Copkalo Nikolaj Nikolaevich – Southern Federal University; e-mail: melnikov@linfotech.ru; Taganrog, Russia; senior researcher; cand. of eng. sc.

УДК 004.75 DOI 10.18522/2311-3103-2021-7-142-153

А.М. Альбертьян, И.И. Курочкин, Э.И. Ватутин

ИСПОЛЬЗОВАНИЕ ГЕТЕРОГЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ УЗЛОВ В ГРИД-СИСТЕМАХ ПРИ РЕШЕНИИ КОМБИНАТОРНЫХ ЗАДАЧ

В настоящее время для решения больших вычислительных задач используются не только многопроцессорные вычислительные системы, но и различные виды распределенных систем. Распределенные вычислительные системы имеют ряд особенностей: возможное наличие отказов узлов и каналов связи, непостоянное время работы узлов, возможные ошибки в расчетах, гетерогенность вычислительных узлов. Под гетерогенностью вычислительных узлов будем понимать не только различную вычислительную способность и различные архитектуры центральных процессоров, но и наличие на узле других компонентов, способных проводить вычисления. К таким компонентам можно отнести видеокарты и математические сопроцессоры. Узел распределенной вычислительной системы будем называть гетерогенным, если помимо одного или нескольких центральных процессоров в его составе есть дополнительные вычислительные устройства. При решении вычислительной задачи на распределенной системе необходимо максимизировать использование всех доступных вычислительных ресурсов. Для этого необходимо не только распределить вычислительные подзадачи на узлы в соответствии с их вычислительной способностью, но и учесть особенности дополнительных вычислительных устройств. Исследованию методов максимизации использования ресурсов на гетерогенных узлах распределенной вычислительной системы посвящена эта работа. Основной целью данной работы является создание переносимого приложения, производящего параллельные вычисления с использованием многопоточной модели выполнения. При разработке приложения акцент делается на наиболее полном использовании доступных аппаратных ресурсов. Одним из основных требований к реализации является оптимизация производительности приложения для различных компьютерных архитектур, а также возможность параллельного выполнения приложения на разнородных вычислительных устройствах, входящих в состав гетерогенного вычислительного комплекса. Была исследована возможность применения ряда методов программно-алгоритмической оптимизации для многопроцессорных архитектур различных поколений. А также была проведена оценка эффективности их использования для высоконагруженных многопоточных приложений. Представлено решение проблемы квазиоптимального динамического распределения вычислительных заданий между всеми доступными на данный момент вычислительными устройствами гетерогенного вычислительного комплекса.

Распределенные вычисления; многопоточное приложение; гетерогенный вычислительный комплекс; повышение производительности; распределение вычислительных ресурсов; грид-система из персональных компьютеров; сопроцессор; Xeon Phi; ортогональные диагональные латинские квадраты; ДЛК.