Раздел II. Математическое и системное программное обеспечение суперкомпьютеров

УДК 004.931 DOI 10.18522/2311-3103-2020-7-35-45

Д.В. Вахлаков, С.Ю. Мельников, В.А. Пересыпкин

МНОГОЭТАПНЫЙ МЕТОД АВТОМАТИЧЕСКОЙ КОРРЕКЦИИ ИСКАЖЕННЫХ ТЕКСТОВ*

Одним из основных факторов, существенно затрудняющих понимание, перевод и анализ текстов, полученных при автоматическом распознавании речи или оптическом распознавании изображений текстов, являются содержащиеся в них искажения в виде ошибочных символов, слов и словосочетаний. 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. Автоматическая обработка таких текстов весьма затруднительна, хотя эта задача является актуальной как для русского, так и для других распространенных языков. Программные средства коррекции, хорошо работающие при малых искажениях в тексте, в случае текстов с высоким уровнем искажений, вне зависимости от их происхождения, показывают неудовлетворительные результаты. Это делает необходимым разработку самостоятельных подходов к коррекции искаженных текстов. Предложен новый многоэтапный метод коррекции искаженных текстов, основанный на последовательном определении ошибок и исправлении искаженных текстов. Искаженными считаются несловарные словоформы и словоформы, вероятность появления которых в тексте в соответствии с выбранной вероятностной моделью меньше заданного порога. 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. 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. Метод позволяет заметно повысить качество (точность) коррекции. В проведенных экспериментах качество коррекции в терминах F1-меры для средне искаженных текстов повысилось на 9 %, а для сильно искаженных текстов – на 7.7 %.

Коррекция текста; искаженный текст; исправление ошибок в тексте; модель языка; расстояние Левенштейна; F1-мера.

* Исследование выполнено при частичной финансовой поддержке РФФИ в рамках научного проекта № 18-29-22104.

D.V. Vakhlakov, S.Yu. Melnikov, V.A. Peresypkin

MULTI-PASS METHOD FOR AUTOMATIC CORRECTION OF DISTORTED TEXTS

One of the main factors that significantly complicate the understanding, translation and analysis of texts obtained by automatic speech recognition or optical recognition of text images are the distortions contained in them in the form of erroneous characters, words and phrases. The most typical errors of recognition systems are: – replacement of a word with a similar sounding or graphic spelling; – replacing several words with one; – replacement of one word with several; – skipping words; – insertion or deletion of short words (including prepositions and conjunctions). As a result of recognition, a text is obtained that has distortions and consists mainly of dictionary words, including in places of distortion. With a large amount of distortion, the texts become almost unreadable. Automatic processing of such texts is very difficult, although this task is relevant both for Russian and for other common languages. Correction software that works well at low distortions in the text, in the case of texts with a high level of distortion, regardless of their origin, show unsatisfactory results. This makes it necessary to develop independent approaches to correcting distorted texts. A new multi-pass method for correction of distorted texts based on sequential error identification and correction of distorted texts is proposed. Non-dictionary word forms and word forms which occurrence probability in the text in accordance with the selected probabilistic model is less than a preset threshold are considered to be distorted. After setting of the distortion sign for individual words, this sign is spread to their combinations, i.e. distorted text fragments are extracted. A list of possible word variants which includes only those word forms from the dictionary that are located at a certain Levenshtein distance from the word under study is built for them. The corrected text from word variants is obtained by searching for the most probable chain of word forms. The correction method consists of several passes, at each pass only those fragments of the text are corrected that remained distorted after the previous pass of correction. The method allows to increase significantly the quality (accuracy) of the correction. In the carried out experiments the quality of correction in terms of the F1-measure for moderately distorted texts has been increased by 9 %, and for highly distorted texts – by 7.7 %.

Language model; automatic text correction; distorted text; noisy text; F1-measure; Levenshtein distance.

Введение. Актуальность задачи. Одним из основных факторов, существенно затрудняющих понимание, перевод и анализ текстов, полученных при автоматическом распознавании речи или оптическом распознавании изображений текстов, являются содержащиеся в них искажения в виде ошибочных символов, слов и словосочетаний ([1–3]). При большом количестве искажений тексты становятся практически нечитаемыми. Автоматическая обработка таких текстов весьма затруднительна, хотя эта задача является актуальной как для русского, так и для других распространенных языков ([4]).

В [5] показано, что распространенные программные средства коррекции ([6–9]), хорошо работающие при малых искажениях в тексте, в случае текстов с высоким уровнем искажений, вне зависимости от их происхождения (набранных с ошибками на клавиатуре, полученных в результате распознавания речи в условиях шумов и др.), показывают неудовлетворительные результаты. Это делает необходимым разработку самостоятельных подходов к коррекции сильно искаженных текстов.

Задача коррекции искаженных текстов, полученных теми или иными системами распознавания, в последние годы формируется как отдельное направление (пост-обработка) и привлекает значительное внимание исследователей. Так, в 2017 и 2019 гг. в рамках конференции International Conference on Document Analysis and Recognition (ICDAR) проводились соревнования различных систем коррекции текстов, полученных в результате оптического распознавания [10, 11]. В соревнованиях принимало участие более 30 участников, и если в 2017 году рассматривались

тексты на двух языках (английский и французский), то в соревнованиях 2019 года к ним добавились болгарский, чешский, нидерландский, финский, немецкий, польский, испанский и словацкий языки.

Отметим, что близкие к описанным постановки рассматриваются также в биоинформационных задачах секвенирования и сборки больших геномов [12] для коррекции так называемых чтений, получаемых с помощью машин-секвенаторов. Эти чтения могут содержать ошибки, поскольку секвенирование основано на выполнении ряда химических реакций, исправление ошибок в наборе чтений является одним из необходимых этапов сборки генома.

В большинстве исследований (см. обзор в [13]) для коррекции используются статистики сочетаемости слов в текстах, т.н. вероятностные языковые модели, которые позволяют строить цепочки словоформ (слов) скорректированного текста из колонок вариантов слов для каждого искаженного фрагмента текста ([14]). В настоящей работе предложен новый многоэтапный метод автоматической коррекции искаженных текстов, обоснованы его преимущества с позиций точности распознавания и скорости работы на современных вычислительных средствах.

Многоэтапные методы коррекции и близкие подходы. Идеи многоэтапных алгоритмов коррекции ошибок использовались разными авторами, прежде всего при разработке систем распознавания речи. В [15] описано применение сложных лингвистических моделей для распознавания при сохранении баланса между сложностью и эффективностью. Предложенная структура состоит из трех этапов: начальное распознавание, обнаружение ошибок и исправление ошибок. Представлен и оценен прототип трехэтапного метода распознавания диктовки на мандаринском диалекте. В этом прототипе первый проход распознает речь с помощью хорошо обученного распознавателя, включающего эффективную языковую модель; второй проход обнаруживает ошибки распознавания с помощью процедуры обнаружения ошибок; третий проход исправляет ошибки, обнаруженные в полученных слабо искаженных текстах. Алгоритм исправления ошибок исправляет ошибки распознавания, сначала создавая списки кандидатов для ошибок, а затем повторно ранжируя кандидатов с помощью триграммной языковой модели.

Для построения множества слов–кандидатов на замену ошибочного, помимо расстояния Левенштейна, могут применяться также его модификации [16], позволяющие точнее корректировать случаи ошибочного разбиения слова на несколько или ошибочного соединения (склейки) рядом стоящих слов.

В работах [17] и [18] предложены усовершенствования этапа коррекции ошибок. Наряду с используемой вероятностной моделью текста, вводится эвристическая мера уверенности, названная «жизнеспособностью слова», которая позволяет исправлять некоторые синтаксические и семантические ошибки за счет учета информации о соседних словах. Идея метода состоит в переупорядочивании списка слов-кандидатов на замену ошибочного. Алгоритм исправления ошибок исправляет ошибки распознавания, сначала создавая список кандидатов для ошибок, а затем повторно ранжируя кандидатов с помощью комбинации оценки взаимной информации, модельной вероятности триграммы и введенной меры «жизнеспособности слова».

Выбор слов-кандидатов на замену ошибочного слова достаточно трудоемок, что связано с многочисленными вычислениями по языковой модели. В [19] предложен способ, ориентированный на снижение вычислительной трудоемкости процедуры исправления ошибок. При выборе кандидатов на замену используются не только слова, близкие по Левенштейну, но и слова, построенные из решеток распознавания, сгенерированных во время распознавания речи. Используется расширенная языковая модель на триграммах слов, которая учитывает их взаимную информацию, а также факторная модель языка, построенная на частях речи.

В работе [20] для повышения точности систем распознавания речи предложен метод исправления ошибок, использующий расширенные знания о предметной области распознаваемого текста. При ранжировании списка слов-кандидатов используется расстояние фонетического редактирования для выбора фонетически близких кандидатов, а для поиска окончательного результата применяется тематически-ориентированная языковая модель.

В [21] предложен подход к обнаружению ошибок в тексте после распознавания речи, основанный на машинном обучении. Для обучения используется большое число признаков, как фонетических, так и связанных со строением предложений текста.

В [22] используется идея привлечения внешних ресурсов для исправления ошибок. Результат распознавания речи проверяется с помощью программной подсказки правописания Bing для обнаружения и исправления неправильно распознанных слов. Текст разбивается на отрезки, состоящие из нескольких слов, и эти отрезки отправляются в качестве поисковых запросов в программную систему Bing. Возвращенный вариант написания означает, что запрос написан с ошибкой, он заменяется предлагаемым исправлением; в противном случае коррекция не выполняется, и алгоритм переходит к следующему отрезку до тех пор, пока не будут проверен и исправлен весь распознанный текст. Похожий подход предложен в [23]. Орфографические ошибки в словах текста, полученного в результате распознавания речи, корректируются на основе набора данных Microsoft N-Gram.

Многоэтапные алгоритмы обработки также используются в области оптического распознавания в задачах сегментации изображений на текстовые блоки ([24, 25]).

Коррекция искаженных текстов с одновременным поиском ошибок. Наиболее характерными ошибками систем распознавания речи и изображений текстов 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. В качестве единицы искаженного текста обычно рассматривается слово.

Если в тексте имеется слово, не входящее в словарь словоформ языка, либо имеющее низкую вероятность в соответствии с выбранной вероятностно-лингвистической моделью языка (как правило, это *N*-граммная модель Маркова на словах), то оно помечается как искаженное, а в качестве возможных вариантов его исправления используются слова из словаря, характеризующиеся высокой степенью сходства с ошибочным словом по некоторой мере. Меру сходства корректируемого слова и слова из словаря вычислятся с использованием расстояния Левенштейна, которое равно минимальному количеству изменений (вставок, замен и удалений) символов алфавита, необходимых для преобразования одного слова в другое.

Таким образом, для каждого слова исходного искаженного текста T подбираются возможные варианты близких по расстоянию Левенштейна слов с помощью их выбора из словаря словоформ языка. В получившихся колонках вариантов слов строятся всевозможные последовательности словоформ (цепочки слов) $S = s_1, s_2, ..., s_n$ и надо найти такую цепочку слов $\hat{S}$, для которой достигается максимум условной вероятности $P(S/T)$:

$$\hat{S} = \max_S P(S/T).$$

В соответствии с формулой Байеса справедливо равенство

$$\max_S P(S/T) = \max_S (P(S) \bullet P(T/S)). \quad (1)$$

Вероятность $P(S)$ определяется моделью языка. В частности, в рамках N-граммной словарной модели эта вероятность представима в виде произведения N условных вероятностей: $P(S) = \prod_{i=1}^{n} P(s_i / s_{i-N+1}, s_{i-N+2}, \ldots s_{i-1})$.

Условная вероятность $P(T/S)$ определяется распределением случайных искажений, которым подвергается текст. Если эти искажения возникают в результате процедуры распознавания, получить необходимые для вычислений по формуле (1) значения $P(T/S)$ практически невозможно. В этом случае можно использовать модели случайных искажений текста (в частности, анализировавшиеся в [26] и [5]).

Для вычисления максимума правой части формулы (1) используются те или иные алгоритмы дискретной оптимизации, например, алгоритмы динамического программирования.

Несмотря на адекватность описанного подхода к коррекции ошибок, он имеет недостатки: используемые на практике модели языка и модели искажений (особенно для систем распознавания) являются достаточно грубыми; алгоритмы дискретной оптимизации в условиях таких сложных оптимизируемых функций чрезвычайно трудоемки и могут обеспечивать поиск лишь локальных оптимумов.

Возможным компромиссом являются многоэтапные методы, обеспечивающие последовательные приближения к оптимальному решению.

Описание нового многоэтапного метода автоматической коррекции искаженных текстов. Предлагаемый метод основан на многоэтапном применении описанного выше подхода, причем на каждом этапе корректируются лишь те фрагменты текста, которые остались искаженными после предыдущего этапа коррекции.

Искаженными считаются несловарные словоформы и словоформы, вероятность появления которых в тексте в соответствии с выбранной вероятностной моделью меньше заданного порога. Словоформы определяются как непрерывные последовательности буквенных символов, отделённых друг от друга пробелами или знаками препинания.

После установки признака искаженности для отдельных слов происходит распространение этого признака на их сочетания, т.е. выделяются искаженные фрагменты текста (ИФТ). Опишем подробнее подход к определению ИФТ.

1. Искаженное слово *А* – это ИФТ.

2. Если слова *А* и *Б* являются ИФТ и между ними присутствует пробел или несколько пробелов, то конкатенация «*А Б*» – тоже ИФТ. Пример объединения участков приведен на рис. 1.

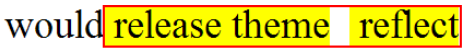

Рис. 1. Пример объединения двух ИФТ, между которыми находится знак пробела

3. Если участки *А* и *Б* являются ИФТ и между ними присутствует слово, в котором менее *d* символов (*d* – параметр), то конкатенация *«А W Б»* – тоже ИФТ (рис. 2).

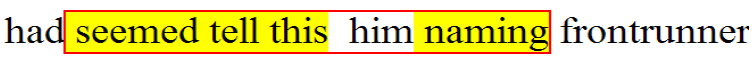

Рис. 2. Пример объединения двух ИФТ, между которыми слово из 3-х символов (для значения d>3)

Опишем предложенный метод коррекции искаженных текстов в виде *k*-этапной процедуры.

На вход поступает искаженный текст, модель языка и словарь словоформ языка.

В качестве параметров метода выступают: k – количество этапов, d – количество символов при определении ИФТ.

1 этап. В исходном тексте с искажениями после определения ИФТ для каждого входящего в них слова строится список возможных вариантов слов, в который попадают только те словоформы из словаря, которые находятся от исследуемого слова на расстоянии Левенштейна, равном 1. Скорректированный текст из вариантов слов получается в результате поиска наиболее вероятной цепочки словоформ в соответствии с (1).

2 этап. В скорректированном после 1 этапа тексте снова определяются ИФТ; для всех искаженных слов, входящих в них, строятся новые списки кандидатов, куда попадают слова, находящиеся на расстоянии Левенштейна, равном 2. Далее в построенных списках ищется наиболее вероятная цепочка словоформ, которая является скорректированным текстом.

…..

k-й этап. В скорректированном на (*k-1*)-м этапе тексте определяются ИФТ, для входящих в них слов строятся списки, куда попадают слова, находящиеся от искаженных слов на расстоянии Левенштейна, равном k. В построенных списках ищется наиболее вероятная цепочка словоформ.

Результатом работы является выход последнего этапа.

Таким образом, алгоритм коррекции состоит из нескольких этапов, на каждом из которых используется определенное значение расстояния Левенштейна, а входными данными являются результаты предыдущего этапа.

Описание и результаты экспериментов. Эксперименты по коррекции проводились с двумя группами текстов, со средними (70 текстов) и сильными искажениями (50 текстов). Процент искаженных слов в текстах первой группы составлял от 15 до 35%, в текстах второй группы от 36 до 50%. Длина текстов варьировалась в пределах от 3000 до 5000 символов. Искажения проводились с помощью программной процедуры, описанной в [5], и являлись комбинацией случайных словарных и символьных искажений.

В качестве языковых моделей использовались 4-граммные модели с модифицированным сглаживанием Кнессера-Нея, построенные на корпусе объемом 200 млн. слов.

Программная реализация описанного подхода тестировалась на вычислителе следующей конфигурации: процессор Intel (R) Xeon (R) CPU E5-2699 v4 @ 2.20GHz, 44 ядра, ОЗУ 250 ГБ, с установленной ОС Windows Server 2012 R2 Standard, x6.

Целью экспериментов являлось получение оценок меры качества и скорости коррекции при различных параметрах алгоритма.

Мерой качества коррекции служила $F1$-мера, которая вычисляется как гармоническое среднее точности A и полноты R коррекций искажённого текста с одинаковым весом, т.е.

$$F1 = \frac{2AR}{(A+R)}. \qquad (2)$$

Полнота коррекции R рассчитывалась как отношение количества верно скорректированных слов $W(T)$ к количеству слов в искажённых фрагментах $W(E)$,

$$R = \frac{W(T)}{W(E)}. \quad (3)$$

а точность коррекции A – через отношение количества слов $W(F)$ в неверных коррекциях к количеству слов в искажённых фрагментах $W(E)$,

$$A = 1 - \frac{W(F)}{W(E)}. \quad (4)$$

В табл. 1 представлены экспериментальные данные по качеству ($F1$-мера (2)) и скорости коррекции (количество скорректированных слов в секунду) программной реализации предложенного подхода.

Таблица 1

Качество и скорость коррекции

	Тексты 1 группы	Тексты 2 группы	Все тексты
Коррекция одноэтапным методом			
Качество ($F1$-мера)	69.3	54.3	63.1
Скорость (слов/сек)	9414.5	895.3	5864.8
Коррекция 4-х этапным методом			
Качество ($F1$-мера)	75.5	58.5	68.4
Скорость (слов/сек)	2007.8	374.4	1327.2

На рис. 3 и 4 представлены графики распределения значений $F1$-меры при коррекции средне и сильно искаженных текстов. Приведены отдельные графики для различного числа этапов (k = 1, 2, 3, 4) многоэтапного метода. Для одноэтапного варианта метода также приведены графики распределения $F1$-меры в четырех случаях, когда список слов-кандидатов составлялся из слов, находящихся на расстоянии Левенштейна L =1, 2, 3, 4 от корректируемого слова.

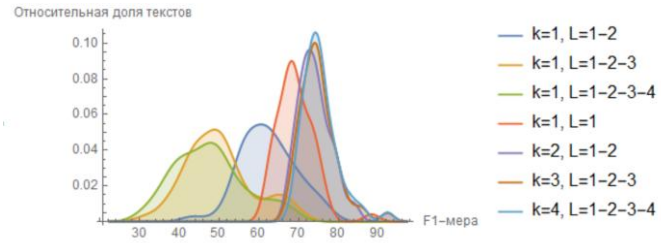

Рис. 3. Распределение значений $F1$-меры после коррекции средне искаженных текстов

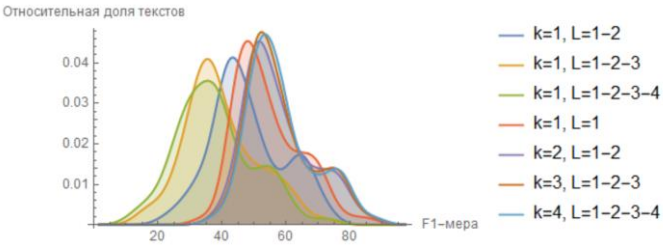

Рис. 4. Распределение значений $F1$-меры после коррекции сильно искаженных текстов

Выводы. Предложен новый многоэтапный метод коррекции искаженных текстов, основанный на последовательном определении ошибок и исправлении искаженных текстов.

Метод позволяет заметно повысить точность коррекции. В проведенных экспериментах качество коррекции в терминах F1-меры для средне искаженных текстов повысилось на 9 %, а для сильно искаженных текстов – на 7.7 %.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Петрова О. О., Булатов К. Б.* Методы пост-обработки результатов распознавания машиночитаемой зоны документов // Тр. ИСА РАН. Специальный выпуск. – 2018. –С. 43-50.
2. *Lee C., Wu S., Liu C., Lee H.* Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension // Proc. Interspeech. – 2018. – P. 3459-3463.
3. *Мещеряков Р.В.* Структура систем синтеза и распознавания речи // Известия Томского политехн. ун-та. – 2009. – Т. 315, № 5. – С. 127-132.
4. *Шакиров И.Ш., Калаков Б.А.* Автоматизация ручной корректировки ошибок оптического распознавания символов // Инженерные решения. – 2020. – № 3 (13). – С. 7-13.
5. *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 (202). – С. 104-114.
6. Спеллер – Технологии Яндекса. – URL: https://tech.yandex.ru/speller/ (дата обращения: 08.11.2020).
7. AfterScan – post-OCR text proofing, advanced spell-checking, automatic correction. – URL: http://www.afterscan.com/ru/ (accessed: 08.11.2020).
8. *Турдаков Д. и др.* Texterra: инфраструктура для анализа текстов // Тр. Института системного программирования РАН. – 2014. – Т. 26. – Вып. 1. – С. 421-438.
9. Microsoft Cognitive Services – API Bing проверки орфографии. – URL: https://www.microsoft.com/en-us/bing/apis/bing-spell-check-api (accessed: 08.11.2020).
10. *Chiron G., Doucet A., Coustaty M., Moreux J.P.* ICDAR 2017 competition on post-OCR text correction // 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). – 2017. – Vol. 1. – P. 1423-1428.
11. *Rigaud C., Doucet A., Coustaty M., Moreux J.P.* ICDAR 2019 Competition on Post-OCR Text Correction // International Conference on Document Analysis and Recognition. – 2019. – P. 1588-1593.
12. *Das A.K., Goswami S., Lee K., Park S.J.* A hybrid and scalable error correction algorithm for indel and substitution errors of long reads // BMC Genomics. – 2019. – Vol. 20 (Suppl 11). – P. 1-15.
13. *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.
14. *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.
15. *Zhou Z., Meng H., Lo W.* A multi-pass error detection and correction framework for Mandarin LVCSR // In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). – 2006. – P. 1646–1649.
16. *Nguyen T.-T.-H., Coustaty M., Doucet A., Jatowt A., Nguyen N.-V.* Adaptive Edit-Distance and Regression Approach for Post-OCR Text Correction / In: Dobreva M., Hinze A., Žumer M. (eds) // Maturity and Innovation in Digital Libraries. ICADL 2018: Lecture Notes in Computer Science. – Vol. 11279. – P. 278-289.
17. *Zukerman I., Partovi A.* Improving the understanding of spoken referring expressions through syntactic-semantic and contextual-phonetic error-correction // Computer Speech & Language. – 2017. – Vol. 46. – P. 284-310.
18. *Li B., Chang F., Liu G.* Speech Recognition error correction by using combinational measures // 3rd IEEE International Conference on Network Infrastructure and Digital Content, Beijing, 2012. – P. 375-379.

19. *Zhou Z.* An error detection and correction framework to improve large vocabulary continuous speech recognition. PhD Thesis, HK, 2009.
20. *Ning Y., Xing C., Zhang L.* Domain Knowledge Enhanced Error Correction Service for Intelligent Speech Interaction / In: Wang D., Zhang LJ. (eds) // Artificial Intelligence and Mobile Services – AIMS 2019: Lecture Notes in Computer Science. – Vol. 11516. – P. 179-187.
21. *Zavareh F., Zukerman I., Kim S., Kleinbauer T.* Error Detection in Automatic Speech Recognition // In Proceedings of Australasian Language Technology Association Workshop. – 2013. – P. 101-105.
22. *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.
23. *Bassil Y., Semaan P.* ASR Context-Sensitive Error Correction Based on Microsoft N-Gram Dataset // Journal of Computing. – January 2012. – Vol. 4, I.1. – P. 34-42.
24. *Abuhaiba I.* Skew Correction of Textural Documents // Journal of King Saud University – Computer and Information Sciences. – 2003. – Vol. 15. – P. 73-93.
25. *Cao H., Prasad R., Natarajan P., MacRostie E.* Robust page segmentation based on smearing and error correction unifying top-down and bottom-up approaches // In: Ninth Internat. Conf. on Document Analysis and Recognition, Curitiba, Brazil, 2007. – P. 392-396.
26. *Белозеров А.А., Вахлаков Д.В., Мельников С.Ю., Пересыпкин В.А., Скавинская Д.В.* Использование эволюционных методов дискретной оптимизации для коррекции искаженных текстов // Вестник компьютерных и информационных технологий. – 2018. – № 12. – С. 3-10.

REFERNCES

1. *Petrova O. O., Bulatov K. B.* Metody post-obrabotki rezul'tatov raspoznavaniya mashinochitaemoy zony dokumentov [Methods of post-processing of results of recognition of the machine-readable zone of documents], *Tr. ISA RAN. Spetsial'nyy vypusk* [Proceedings of the ISA RAS. Special issue], 2018, pp. 43-50.
2. *Lee C., Wu S., Liu C., Lee H.* Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension, *Proc. Interspeech.*, 2018, pp. 3459-3463.
3. *Meshcheryakov R.V.* Struktura sistem sinteza i raspoznavaniya rechi [Structure of systems synthesis and speech recognition], *Izvestiya Tomskogo politekhn. un-ta* [Proceedings of the Tomsk Polytechnic University], 2009, Vol. 315, No. 5, pp. 127-132.
4. *Shakirov I.Sh., Kalakov B.A.* Avtomatizatsiya ruchnoy korrektirovki oshibok opticheskogo raspoznavaniya simvolov [Automating manual correction of optical character recognition errors], *Inzhenernye resheniya* [Engineering solutions], 2020, No. 3 (13), pp. 7-13.
5. *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 correction tools for distorted texts depending on the nature of the distortion], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2018, No. 8 (202), pp. 104-114.
6. Speller – Tekhnologii Yandeksa [Speller-Yandex Technologies]. Available at: https://tech.yandex.ru/speller/ (accessed 08 November 2020).
7. AfterScan – post-OCR text proofing, advanced spell-checking, automatic correction. Available at: http://www.afterscan.com/ru/ (accessed 08 November 2020).
8. *Turdakov D. i dr.* Texterra: infrastruktura dlya analiza tekstov [Texterra: infrastructure for text analysis], *Tr. Instituta sistemnogo programmirovaniya RAN* [Proceedings of the Institute of System Programming of the Russian Academy of Sciences], 2014, Vol. 26, Issue 1, pp. 421-438.
9. Microsoft Cognitive Services – API Bing проверки орфографии. Available at: https://www.microsoft.com/en-us/bing/apis/bing-spell-check-api (accessed 08 November 2020).
10. *Chiron G., Doucet A., Coustaty M., Moreux J.P.* ICDAR 2017 competition on post-OCR text correction, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, Vol. 1, pp. 1423-1428.
11. *Rigaud C., Doucet A., Coustaty M., Moreux J.P.* ICDAR 2019 Competition on Post-OCR Text Correction, *International Conference on Document Analysis and Recognition*, 2019, pp. 1588-1593.
12. *Das A.K., Goswami S., Lee K., Park S.J.* A hybrid and scalable error correction algorithm for indel and substitution errors of long reads, *BMC Genomics*, 2019. Vol. 20 (Suppl 11), pp. 1-15.

13. *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 the language. Software system for evaluating the readability of distorted texts], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2019, No. 8, pp. 6-18.
14. *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 for detecting errors in distorted texts], *Vestnik komp'yuternykh i informatsionnykh tekhnologiy* [Bulletin of Computer and Information Technologies], 2016, No. 5, pp. 29-34.
15. *Zhou Z., Meng H., Lo W.* A multi-pass error detection and correction framework for Mandarin LVCSR, I*n: Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2006, pp. 1646-1649.
16. *Nguyen T.-T.-H., Coustaty M., Doucet A., Jatowt A., Nguyen N.-V.* Adaptive Edit-Distance and Regression Approach for Post-OCR Text Correction, In: Dobreva M., Hinze A., Žumer M. (eds), *Maturity and Innovation in Digital Libraries. ICADL 2018: Lecture Notes in Computer Science*, Vol. 11279, pp. 278-289.
17. *Zukerman I., Partovi A.* Improving the understanding of spoken referring expressions through syntactic-semantic and contextual-phonetic error-correction, *Computer Speech & Language*, 2017, Vol. 46, pp. 284-310.
18. *Li B., Chang F., Liu G.* Speech Recognition error correction by using combinational measures, *3rd IEEE International Conference on Network Infrastructure and Digital Content, Beijing, 2012*, pp. 375-379.
19. *Zhou Z.* An error detection and correction framework to improve large vocabulary continuous speech recognition. PhD Thesis, HK, 2009.
20. *Ning Y., Xing C., Zhang L.* Domain Knowledge Enhanced Error Correction Service for Intelligent Speech Interaction, In: Wang D., Zhang LJ. (eds), *Artificial Intelligence and Mobile Services – AIMS 2019: Lecture Notes in Computer Science*, Vol. 11516, pp. 179-187.
21. *Zavareh F., Zukerman I., Kim S., Kleinbauer T.* Error Detection in Automatic Speech Recognition, *In Proceedings of Australasian Language Technology Association Workshop*, 2013, pp. 101-105.
22. *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.
23. *Bassil Y., Semaan P.* ASR Context-Sensitive Error Correction Based on Microsoft N-Gram Dataset, *Journal of Computing*, January 2012, Vol. 4, I.1, pp. 34-42.
24. *Abuhaiba I.* Skew Correction of Textural Documents, *Journal of King Saud University – Computer and Information Sciences*, 2003, Vol. 15, pp. 73-93.
25. *Cao H., Prasad R., Natarajan P., MacRostie E.* Robust page segmentation based on smearing and error correction unifying top-down and bottom-up approaches, *In: Ninth Internat. Conf. on Document Analysis and Recognition, Curitiba, Brazil, 2007*, pp. 392-396.
26. *Belozerov A.A., Vakhlakov D.V., Mel'nikov S.Yu., Peresypkin V.A., Skavinskaya D.V.* Ispol'zovanie evolyutsionnykh metodov diskretnoy optimizatsii dlya korrektsii iskazhennykh tekstov [The use of evolutionary methods of discrete optimization for correction of distorted texts], *Vestnik komp'yuternykh i informatsionnykh tekhnologiy* [Bulletin of Computer and Information Technologies], 2018, No. 12, pp. 3-10.

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

Вахлаков Дмитрий Владимирович – ФГУП «НТЦ «Орион»; e-mail: melnikov@linfotech.ru; 127018, г. Москва, ул. Образцова, д. 38, стр. 1; тел/факс: +74952499053.

Пересыпкин Владимир Анатольевич – к.т.н.

Мельников Сергей Юрьевич – ООО «Линфо»; e-mail: melnikov@linfotech.ru; 127018, г. Москва, ул. Образцова, д. 38, стр. 1; тел/факс: +74952499053, моб.: +79037222824; к.ф.-м.н.; инженер-математик.

Vakhlakov Dmitriy Vladimirivich – FGUP "NTC "Orion"; e-mail: melnikov@linfotech.ru; 38, Obraztsova street, build. 1, Moscow, 127018, Russia; phone/fax: +74952499053.

Peresypkin Vladimir Anatolyevich – cand. of eng. sc.

Melnikov Sergey Yurievich – Linfo LLC; e-mail: melnikov@linfotech.ru; 38, Obraztsova street, build. 1, Moscow, 127018, Russia; phone/fax: +74952499053, mobile: +79037222824; cand. of eng. sc.; mathematical engineer.

УДК 004.056.55 DOI 10.18522/2311-3103-2020-7-45-52

Е.И. Духнич, А.Г. Чефранов

АППАРАТУРНО-ОРИЕНТИРОВАННЫЙ АЛГОРИТМ ДЛЯ БЫСТРОГО УМНОЖЕНИЯ КРОНЕКЕРОВА ПРОИЗВЕДЕНИЯ МАТРИЦ НА ВЕКТОР

В статье на основе использования свойств произведения Кронекера (КП) матриц предлагается новый алгоритм для повышения эффективности выполнения операции умножения КП на вектор. Указанная операция широко применяется при решении задач обработки сигналов, изображений, криптографии и т.п., где выполняется формирование матриц большого размера с заданными свойствами с помощью КП матриц малого размера. При этом используются матрицы со следующими свойствами: ортогональные (унитарные), обращаемые, инволютивные. Умножение квадратной матрицы размера $n \times n$ на вектор имеет вычислительную сложность $O(n^2)$. Поэтому при росте количества элементарных матриц-сомножителей размер результирующей матрицы КП и сложность умножения ее на вектор растут экспоненциально. Это обстоятельство существенно повышает время решения прикладных задач. Целью предлагаемой работы является построение алгоритма, ориентированного на аппаратную реализацию и ускоряющего процессы формирования КП и умножения вектора на него. Предлагается совместить во времени эти процедуры. Таким образом матрица КП в явном виде фактически не рассчитывается. Вместо этого матрицы-сомножители КП итеративно умножаются на компоненты вектора за время $O(n\log_2 n)$ и требуют линейной сложности памяти. Приведена схема вычислений с топологией гиперкуба для возможной аппаратной реализации предлагаемого алгоритма, которая легко поддается конвейеризации. В разделе 1 приведены определения и свойства КП, используемые при синтезе предлагаемого алгоритма. В разделе 2 рассмотрен иллюстрирующий предлагаемый алгоритм пример с $n = 8$, на основе которого в разделе 3 предложена аппаратурно-ориентированная структура его реализации для произвольного n.

Алгоритм, произведение Кронекера; элементарная матрица; сложность вычислений; конвейерная реализация.

E.I. Dukhnich, A.G. Chefranov

HARDWARE-ORIENTED ALGORITHM FOR FAST MULTIPLICATION OF A VECTOR BY A MATRIX KRONECKER PRODUCT

The article discusses new algorithm to increase the efficiency of the operation of multiplying a matrix Kronecker product (KP) by a vector. It is based on the use of the KP properties. This operation is widely used in solving problems of processing signals, images, cryptography, etc., where the formation of large matrices with specified properties is performed using small size matrices. In this case, matrices with the following properties are used: orthogonal (unitary), invertible, involutive. Multiplying an $n \times n$ square matrix by a vector has a computational complexity of $O(n^2)$. Therefore, with an increase in the number of elementary matrix factors, the size of the resulting KP matrix and the complexity of multiplying it by a vector grow exponentially. This circumstance significantly increases the time for solving applied problems. The aim of the proposed work is to construct an algorithm that accelerates the processes of forming the KP and multiplying