УДК: 519.16, 519.17
DOI: 10.25559/SITITO.16.202002.304-313

О числе реконструкций по подсловам в бинарном алфавите при наложении на один символ

Г. Н. Жукова[1*], **Ю. Г. Сметанин**[2], **М. В. Ульянов**[3,4]
[1] ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия
101000, Россия, г. Москва, ул. Мясницкая, д. 20
[*] gzhukova@hse.ru
[2] ФГУ «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия
119333, Россия, г. Москва, ул. Вавилова, д. 44, корп. 2
[3] ФГБУН «Институт проблем управления им. В.А. Трапезникова Российской академии наук», г. Москва, Россия
117997, Россия, г. Москва, ул. Профсоюзная, д. 65
[4] ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова», г. Москва, Россия
119991, Россия, г. Москва, ГСП-1, Ленинские горы, д. 1

Аннотация

Рассмотрена задача получения точных оценок числа реконструкций для слов над бинарным алфавитом. Подслова различной длины из заданного множества соединяются методом наложения концевых символов. Соединять можно только такую пару подслов, у которых последний символ первого подслова совпадает с первым символом второго. При наложении пары подходящих подслов на один символ из двух одинаковых символов (в конце первого и в начале второго подслова) в реконструкцию входит только один. Предложен подход, основанный на рассмотрении усеченных подслов, состоящих из префикса и суффикса данного подслова, имеющих длину один. При построении реконструкции вместо самих подслов из заданного множества соединяются усеченные слова вида «00», «01», «10» и «11». Число реконструкций находится в предположении, что каждое из усеченных подслов соответствует уникальному подслову в заданном множестве подслов. В результате при соединении слов «00» и «00» возможны две реконструкции, соответствующие соединению исходных подслов «0x0» и «0y0» в «0x0y0» и «0y0x0», где x и y — различные последовательности символов бинарного алфавита, одна из которых может быть пустой (но не обе одновременно).
Такой подход позволил определить условия существования реконструкции по заданному множеству подслов различной длины. Показано, при каких условиях, касающихся количества усеченных подслов каждого вида, реконструкция невозможна. Например, невозможна реконструкция по множеству подслов, содержащему только подслова вида «00» и «11». Также невозможно соединить все подслова заданного множества, если число усеченных подслов вида «01» и «10» отличается больше, чем на один. Для различных случаев, допускающих полную реконструкцию, получены формулы точного числа реконструкций. Точное число реконструкций зависит от наличия или отсутствия подслов, соответствующих усеченным подсловам каждого вида.
Поскольку возможность реконструкции главным образом зависит от соотношения числа подслов вида «01» и «10», то отдельно была рассмотрена модель с возможностью инверсий слов. Предполагается, что множество подслов для реконструкции содержит только слова вида вида «00», «01» «00», Часть слов вида «01» записывается в обратном порядке и становится словами вида «10». Если слов «01» было четное число, то в «01» преобразуется половина слов «01», иначе — половина от ближайшего четного числа. В последнем случае из множества подслов вида «01» получается два варианта наборов подслов вида «01» и «10», в одном больше подслов

«01», в другом — «10». Для каждого случая приведены формулы точного числа реконструкций при условии уникальности подслов в заданном множестве, а также несимметричности подслов, порождающих усеченные подслова вида «00» и «11».

Ключевые слова: комбинаторика слов, бинарный алфавит, символьная последовательность, реконструкция последовательности, реконструкция последовательности по ее подпоследовательностям.

Финансирование: исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 19-07-00150 «Методы построения и анализа информативных символьных представлений по фрагментарным данным».

Для цитирования: Жукова, Г. Н. О числе реконструкций по подсловам в бинарном алфавите при наложении на один символ / Г. Н. Жукова, Ю. Г. Сметанин, М. В. Ульянов. – DOI 10.25559/SITITO.16.202002.304-313 // Современные информационные технологии и ИТ-образование. – 2020. – Т. 16, № 2. – С. 304-313.

On the Number of Subword Reconstructions in the Binary Alphabet when Superimposed on one Character

G. N. Zhukova[a*], Yu. G. Smetanin[b], M. V. Ulyanov[c,d]
[a] National Research University Higher School of Economics, Moscow, Russia
20 Myasnitskaya St., Moscow 101000, Russia
[*] gzhukova@hse.ru
[b] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia
44-2 Vavilova St., Moscow 119333, Russia
[c] V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
65 Profsoyuznaya St., Moscow 117997, Russia
[d] Lomonosov Moscow State University, Moscow, Russia
1, Leninskie gory, Moscow 119991, Russia

Abstract

The problem of obtaining accurate estimates of the number of reconstructions for words over a binary alphabet is considered. Subwords of different lengths from a given set are joined by the method of overlapping end characters. It is only possible to connect a pair of subwords where the last character of the first subword is the same as the first character of the second. When superimposing a pair of suitable subwords on one character of two identical ones (at the end of the first and at the beginning of the second subword) only one, is included in the reconstruction. An approach based on combining truncated subwords consisting of the first and last characters of a subword is proposed. When building a reconstruction, instead of the subwords from a given set, truncated words of the form "00", "01", "10" and "11" are connected. The number of reconstructions is under the assumption that each of the truncated subwords corresponds to a unique subword in a given set of subwords. As a result, when combining the words "00" and "00", two reconstructions are possible, corresponding to combining the original subwords "0x0" and "0y0" into "0x0y0" and "0y0x0", where x and y are different sequences of binary alphabet characters, one of which may be empty (but not both simultaneously).
Such an approach made it possible to determine the conditions for the existence of reconstruction from a given set of subwords of various lengths. It is noted under what conditions, concerning the number of truncated subwords of each type, reconstruction is impossible. For example, reconstruction by a set of subwords containing only subwords of the form "00" and "11" is not possible. It is also impossible to combine all the subwords of a given set if the number of truncated subwords of the form "01" and "10" differs by more than one. For various cases allowing for complete reconstruction, formulas of the exact number of reconstructions are obtained. The exact number of reconstructions depends on the presence or absence of subwords corresponding to truncated subwords of each type.
Since the possibility of reconstruction mainly depends on the ratio of the number of subwords of the form "01" and "10", a model with the possibility of word inversions was also considered. It is assumed that the set of subwords for reconstruction contains only words of the form "00", "01" and "11". Some of the words of the form "01" are written in the reverse order and become words of the form "10". If the words "01" were an even number, then half of the words "01" would be converted to "01", otherwise, half of the nearest even number would be. In the latter case, from the set of subwords of the form "01", two variants of the sets of subwords of the form "01" and "10" are obtained, in one, there are more subwords "01", in the other "10". For each case, formulas are given for the exact number of reconstructions, provided that the subwords in the given set are unique, as well as the asymmetry of the subwords generating truncated subwords of the form "00" and "11".

Keywords: combinatorics on words, binary alphabet, symbolic sequence, reconstruction of a sequence, reconstruction of a sequences from their subsequences.

Funding: This paper is based on research carried out with the financial support of the grant of the Russian Scientific Foundation (project No. 19-07-00150).

For citation: Zhukova G.N., Smetanin Yu.G., Ulyanov M.V. On the Number of Subword Reconstructions in the Binary Alphabet when Superimposed on one Character. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2020; 16(2):304-313. DOI: https://doi.org/10.25559/SITITO.16.202002.304-313

Введение

Одним из способов представления информации в рамках задач качественного анализа систем и процессов является символьное кодирование. Поведение объекта или процесса кодируется словом над конечным алфавитом, которое и является объектом дальнейшего исследования. Однако, в некоторых случаях, например, в ряде задач биоинформатики и геномики, а также и в прикладных областях, связанных с анализом данных, распознаванием образов, прогнозированием и т.д., исследователи получают не само слово целиком, а множество подслов, которые являются последовательными фрагментами некоторого слова. Поскольку при этом позиции подслов в исходном слове неизвестны, возникает задача реконструкции — восстановления слова по исходному множеству подслов. Задачи восстановления слов над конечным алфавитом имеют различные постановки, отличающиеся как объемом имеющейся информацией, так и ограничениями на допустимые решения. Обычно эти задачи, как задачи с неполной информацией, являются сложными. По характеру анализируемой информации эти задачи относятся к области комбинаторики слов [1-7]. Хотя в классической комбинаторике слов непосредственно задача реконструкции не рассматривается, однако методы, созданные в рамках комбинаторики слов, полезны при решении задач реконструкции. При возрастающей неопределенности в исходных данных и комбинаторном росте числа возможных реконструкций для построения различных оценок используются вероятностно-статистические методы [8-12], а при достаточно сильных ограничениях известны методы реконструкции за линейное время [13].

Одной из важнейших областей практического применения методов комбинаторики слов является область биомолекулярных моделей и процессов. Более того, активные исследования в этой области были одной из причин, вызвавших появление и быстрое развитие комбинаторики слов как самостоятельной дисциплины. Вопросы применения комбинаторных методов в анализе ДНК рассматривались многими выдающимися учеными уже более полувека назад; достаточно привести в пример работу Г. Гамова [14].

В настоящее время в различных областях математической биологии в качестве инструментов для анализа процессов на молекулярном уровне используются методы дискретной математики, теоретической информатики и топологии, которые подробно описаны, например, в [15]. Методы комбинаторики слов и теории графов эффективно применимы в задачах анализа структуры и расположения молекул и генов, при встраивании ДНК в пространственные графовые модели динамики и кинетики молекулярных взаимодействий. Отметим также, что задача секвенирования геномов по сути является задачей реконструкции слов в условиях сильных ограничений, подразумевающей однозначность реконструкции.

В рамках исследований, связанных с однозначностью реконструкций, отметим работы, которые внесли вклад не только в биоинформатику, но и в саму комбинаторику слов. Характерный пример — работа [16], где исследованы конечные слова над алфавитом пар букв $\Sigma = \left\{ a, \bar{a}, b, \bar{b} \right\}$, где $\bar{a} = a, \bar{b} = b$, а каждое слово $w_1 w_2 \ldots w_t$ отождествляется со своим обратным дополнением $w_t \ldots w_2 w_1$. В [16] определено наименьшее k такое, что каждое слово длины n, составленное из символов Σ, однозначно определяется набором его подслов длины до k.

Примером другой области, в которых возникают задачи комбинаторики слов и реконструкций с зачастую большим числом решений, является анализ естественного языка. В семантическом анализе естественного языка распространены и крайне эффективны методы машинного обучения, но часто возникает проблема недостаточности данных для удовлетворительной работы этих методов, в то время как люди могут угадывать значение слова по малому числу примеров. Для устранения этой проблемы разрабатываются модели, которые должны позволить использовать фрагментарные фоновые знания о контексте [17]. Основная трудность связана с тем, что в задачах обработки естественного языка с использованием средств машинного обучения часто используются представления, в которых каждому слову назначается отдельный вектор, и при этом игнорируется морфология слов. Чтобы преодолеть этот недостаток, в [18] используется представление каждого слова пакетом символьных n-грамм. При этом возникает задача применения n-грамм для решения таких задач, как оценка подобия слов.

Вложения слов в пространство меньшего размера на основе использования информации о распределении, полученной из немаркированных данных, улучшают обобщение по лексическим признакам, однако эффективность встраивания слов для последующих задач ограничена словами из словаря. Разрабатываются подходы к композиционной генерации вложений слов, не входящих в словарь [19].

В [20] для решения задачи обобщения предварительно обученных вложений слов за пределы словарей фиксированного размера без использования дополнительной контекстной информации предложена модель генерации вектора слов на уровне подслов, которая рассматривает слова как пакеты символьных n-грамм. Модель проста, быстро обучаема и обеспечивает хорошие векторы для редких или невидимых слов. Эксперименты показывают, что модель достигает самых современных характеристик в задаче сходства английских слов и в совместном прогнозировании части речи и морфосинтаксических атрибутов на 23 языках, что свидетельствует о ее способности улавливать связь между текстовыми представлениями слов и их вложениями.

Идея встраивания слов на основе подслов была предложена главным образом для решения проблемы со словами, не входящими в словарь. В [21] предложен метод восстановления предварительно обученных вложений слов с использованием информации о подсловах, которая может эффективно представлять большое количество вложений слов в значительно меньшем фиксированном пространстве. Реконструированные вложения на основе подслов могут успешно имитировать хорошо обученные вложения слов в небольшом фиксированном пространстве, одновременно предотвращая ухудшение качества, и могут прогнозировать эффективные вложения слов не из словаря. К настоящему времени разработаны регрессионные модели, объединяющие преимущества методов факторизации матриц и локального контекстного окна и эффективно использующие статистическую информацию при обучении [22], а также модельные архитектуры для вычисления непрерывных векторных представлений слов из очень больших наборов данных в задаче оценки подобия слов [23].

При оценке числа возможных реконструкций возможны два подхода — получение точного числа реконструкций с восстановлением реконструированных слов или оценка числа реконструкций, полученная на основе вероятностных или каких-либо других предположений. Ранее в [24] авторы предложили

решение идеализированной задачи о реконструкции, когда предполагается, что исходный набор слов одинаковой длины является полным набором подслов для некоторого неизвестного слова. При этом предполагалось, что исходные подслова получены при сдвиге окна на один символ по некоторому неизвестного слову. Был также рассмотрен вариант с дополнительными ограничениями на реконструкцию в виде известного набора запрещенных подслов для этого неизвестного слова [25, 26]. В обоих случаях используется аппарат поиска эйлеровых путей на орграфах. Предложенные методы позволяют получить как точное число реконструкций, так и сами реконструированные слова. Даже в таких постановках возможны ситуации, когда решения не единственны, поскольку возможное число реконструкций равно числу эйлеровых путей или циклов в соответствующем мультиорграфе Де-Брейна [24, 25]. Некоторые оценки числа возможных реконструкций, основанные на статистическом подходе, в предположении о равновероятном распределении символов бинарного алфавита, были сделаны авторами ранее[1] в [27].

Ясно, что в более общих случаях число решений задачи реконструкции может быть очень велико, и поэтому методы оценки числа решений в различных условиях постановки задачи представляют интерес, особенно в условиях недостаточно полной информации. Очевидный интерес представляет получение оценки числа решений задачи реконструкции в случае подслов различной длины, что достаточно характерно для многих предметных областей. Разработка точных оценок для слов над бинарным алфавитом и подслов различной длины при сборке с наложением в один символ и составляет предмет настоящей статьи.

Постановка задачи

Пусть имеется множество слов $|$ мощности n, содержащее различные, т.е. уникальные слова не обязательно одинаковой длины $l \geq 2$ в бинарном алфавите $\{0,1\}$. В общем случае слова из множества $|$ содержат в начале и в конце или только нули, или только единицы, или на одном конце ноль, а на другом единицу. Под полной реконструкцией (сборкой) с наложением один по множеству слов $|$ будем далее понимать создание одного «длинного» слова путем соединения всех слов из $|$ в некотором порядке путем их конкатенации с наложением на один символ, если такая реконструкция возможна. Таким образом, полная реконструкция по множеству слов $|$ есть «длинное» слово, полученное соединением всех слов из этого множества. Далее под термином «реконструкция» мы будем понимать полную реконструкцию с наложением один. Пусть W — множество таких возможных «длинных» слов, т.е. возможных реконструкций по множеству $|$. Мы будем рассматривать задачу определения мощности N множества W, т.е. определения числа возможных реконструкций на основе особенностей слов в $|$. Если $W = \varnothing$, то $N = 0$ и реконструкция из $|$ невозможна. Далее рассматриваются две постановки:

— определение числа реконструкций при запрещенных инверсиях слов из $|$.

— определение числа реконструкций при разрешенных инверсиях слов из $|$.

Реконструкция при запрещенных инверсиях

При запрещенных инверсиях мы имеем тот порядок символов 0 и 1 в словах, в котором эти слова представлены в исходном множестве $|$. Конкатенация двух слов наложением на один символ подразумевает, что можно соединять слова по схеме «голова-хвост», если первое (левое) слово заканчивается на тот же символ, с которого начинается второе (правое) слово. Отметим, что два слова в определенной ситуации допускают два варианта соединения. Например, можно соединить слова «01010000111» и «11000010» наложением единиц, в результате мы получим «0101000011110000010», поскольку на месте соединения одинаковые концевые символы накладываются, и в результате из двух одинаковых символов остается один (это и есть соединение с наложением один). Эти же два слова можно соединить в другом порядке, а именно путем наложения нулей «11000010» и «01010000111», результатом будет «110000101010000111». Жирным шрифтом выделены наложенные символы.

При соединении слов не важно, какие последовательности символов алфавита находятся внутри слова, поэтому далее будем рассматривать вместо самих слов укороченные их варианты, состоящие только из символов начала и конца слова. Ограничение в постановке задачи на длину слов $l \geq 2$ всегда гарантирует существование таких укороченных вариантов. Так, вместо слова «01010000111» будем использовать его укороченный вариант «01». Таким образом, возможных укороченных вариантов всего четыре — «00», «01», «10» и «11». Далее мы будем называть эти укороченные варианты словами из $|$. Очевидно, что реконструкции не всегда существуют, например, если во множестве содержатся одновременно только слова «00» и «11», то $N = 0$. С другой стороны, если состоит из двух слов «01» и «10», их можно соединить двумя способами: «01» + «10» = «010» или «10» + «01» = «101», так что реконструкция (если она существует) может не быть единственной. Пусть n_{00} — число слов вида «00», n_{01} — число слов вида «01», n_{10} — число слов вида «10» и n_{11} — число слов вида «11», тогда общее число слов в $|$ — $n = n_{00} + n_{01} + n_{10} + n_{11}$.

Будем рассматривать различные варианты в зависимости от количества слов того или иного вида в исходном множестве $|$. Основной вопрос — это вопрос о том, при каких условиях возможно построить реконструкцию, т.е. осуществить сборку всех слов из множества $|$ в одно слово. Тривиальный случай $|\Omega| = 1$ и $N = 1$ не представляет интереса, поэтому далее $|\Omega| \geq 2$.

Вначале ответим на вопрос об условиях невозможности реконструкций — ответ на него очевиден, $N = 0$, если:

— $n = n_{00} + n_{11}$, $n_{00} \neq 0$, $n_{11} \neq 0$ — в $|$ есть только слова вида «00» и «11»;

— $|n_{01} - n_{10}| \geq 2$ — количество слов типа «01» отличается от количества слов «10» больше чем на один, причем слова одного из данных типов могут вообще отсутствовать: $n_{01} = 0$ или $n_{10} = 0$.

Далее мы будем рассматривать только случаи, допускающие полную реконструкцию, когда множество слов из $|$ обладает следующими свойствами:

[1] Сметанин, Ю. Г. О числе возможных реконструкций слов по подсловам при окне переменного сдвига / Ю. Г. Сметанин, М. В. Ульянов, А. С. Пестова. —DOI 10.17587/it.24.233-238 // Информационные технологии. — 2018. — Т. 24, № 4. — С. 233-238. — URL: https://www.elibrary.ru/item.asp?id=32837169 (дата обращения: 27.05.2020). — Рез. англ.

a) $n = n_{00}$ — в $|$ есть только слова вида «00» и еще может быть одно слово вида «01» или одно слово «10»;
b) $n = n_{11}$ — в $|$ есть только слова вида «11» и еще может быть одно слово вида «01» или одно слово «10»;
c) $n = n_{00} + n_{01} + n_{10} + n_{11}$, $n_{01} = 1, n_{10} = 0$ или $n_{01} = 0, n_{10} = 1$ — в $|$ содержится любое число слов вида «00» и «11» и одно слово вида «01» или «10»;
d) $n = n_{01} + n_{10}, n_{01} \neq 0, n_{10} \neq 0$, и $|n_{01} - n_{10}| \leq 1$ — в $|$ есть только слова вида «01» и «10», и их количество отличается не более чем на один;
e) $n = n_{00} + n_{01} + n_{10} + n_{11}$ и $|n_{01} - n_{10}| \leq 1$ — в $|$ есть слова вида «01» и «10», и их количество отличается не более чем на один, и любое число слов вида «00» и «11.
Для указанных выше случаев рассмотрим решение задачи о числе N возможных реконструкций с наложением один по множеству без инверсий.

3.1. Случаи a) и b) — множество состоит только из слов вида «00» или слов вида «11»

Поскольку в рассматриваемой постановке все слова во множестве уникальны, причем слова вида «00» («11») можно соединять друг с другом в любой последовательности, то число возможных реконструкций N равно количеству перестановок всех слов из $|$. Поскольку мощность множества $|$ равна , то
$N = n_{00}! = n!$ или $N = n_{11}! = n!$ (1)

Если в $|$ есть еще слово «01» или «10», то $N = n_{00}! = (n-1)!$, поскольку слово «01» или «10» может занимать единственное возможное положение в реконструкции.
Уточним, что в постановке с запрещением инверсий слова из $|$ нельзя инвертировать при построении реконструкции, т.е. слова «010110» и «0100» можно соединить «010110» + «0100» = «010110100» или «0100» + «010110» = «010010110», но не «010110» + «0010» или «0100» + «011010».

3.2. Случай c) — множество состоит слов «00» и/или «11» и одного слова «01» или «10»

Пусть в $|$ есть одно слово вида «01». В этом случае слева строится последовательность слов «00» с числом реконструкций $n_{00}!$, далее слово «01», которое является сцепляющим для последовательности слов «00» и «11», после чего строится последовательность слов «11» с числом реконструкций $n_{11}!$. Очевидно, что реконструкции «00» и «11» независимы, тогда
$N = n_{00}! \cdot n_{11}!$. (2)

Заметим, что при наличии одного слова вида «10» ситуация с числом реконструкций не изменяется, но при этом слева строится последовательность слов вида «11», затем слово «10» и справа от него последовательность слов вида «00». Формула (2) справедлива и в случае, если $n_{00} = 0$ или $n_{11} = 0$.

3.3. Случай d) — множество состоит только из слов вида «01» и «10»

В этом случае $n = n_{01} + n_{10}, n_{01} \neq 0, n_{10} \neq 0$, и $|n_{01} - n_{10}| \leq 1$, рассмотрим отдельно ситуации с равным и не равным числом слов вида «01» и «10».
i) $|n_{01} - n_{10}| = 0$ — во множестве одинаковое количество $n_{01} = n_{10} = n/2$ слов вида «01» и «10», тогда их все можно соединить, чередуя слова «01» и «10», причем можно начать как с «01», так и с «10». В подпоследовательности, состоящей из всех слов «01» уникальные слова можно расположить по порядку $n_{01}!$ способами. Аналогично для подпоследовательности, состоящей из всех слов «10» — $n_{10}!$ вариантов, так что всего получается $n_{01}! \cdot n_{10}!$ реконструкций, начинающихся с «01» и столько же начинающихся с «10», итого всего реконструкций
$N = 2 \cdot n_{01}! \cdot n_{10}! = 2 \cdot \left((n/2)!\right)^2$. (3)

ii) Если во множестве слов вида «01» и слов вида «10», то все их можно соединить, чередуя слова «01» и «10», причем реконструкцию можно начать только со слова «10», при этом разных подпоследовательностей, состоящих из всех слов «01» , а последовательностей из слов «10» поэтому всего мы получаем
$N = m! \cdot (m+1)!$ (4)

реконструкций. Для случая, когда состоит из слов вида «10» и слов вида «01» формула аналогична. Для объединения результата обозначим $\alpha = |n_{01} - n_{10}|$, $\alpha = 0, 1$ разность между числом слов вида «01» и числом слов вида «10», тогда слов одного вида , а другого вида $m + \alpha$, и общая формула для числа реконструкций запишется в виде
$N = m! \cdot (m+\alpha)! \cdot (2-\alpha)$. (5)

3.4. Случай e) — множество $|$ состоит из слов всех видов, и количество слов вида «01» и «10» отличается не более чем на один

В этом случае $n = n_{00} + n_{01} + n_{10} + n_{11}$ и $|n_{01} - n_{10}| \leq 1$. Реконструкции по множеству будем строить, располагая вначале в определенном порядке слова вида «01» и «10», получим число вариантов по формуле (5). Затем будем вставлять слова вида «00» между словами «10» и «01». Поскольку число слов «00» во множестве $|$ равно n_{00}, тогда в любое такое место можно вставить от нуля до n_{00} слов вида «00».
Пусть состоит из n_{01} слов вида «01» и слов вида «10», тогда при соединении слов этих двух видов, начиная с «01», образуется мест для вставки слов вида «00». Если слова «01» и «10» соединять, начиная с «10», то мест для вставки «00» будет . В случае, когда состоит из слов вида «01» и $n_{10} = m+1$ слов вида «10», образуется мест для вставки слов вида «00». Такая же ситуация и при наличии слов вида «10» и слов вида «01».
Обозначим $m = \lfloor (n_{01} + n_{10})/2 \rfloor$, тогда слов «00» нужно распределить по r местам, где

$$r = \begin{cases} m, \text{"01" } \delta \text{ "10 } \textit{поровну, соединены начиная с } \text{"10"} \\ m+1, \textit{иначе} \end{cases}. \quad (6)$$

Для каждого из слов «00» независимо от остальных слов «00» допустимо расположение в любом из r мест между словами «10» и «01» или на конце «0» последовательности. Таким образом, мы имеем всего $r^{n_{00}}$ вариантов распределения слов «00» по местам. Для каждого из таких вариантов распределения уникальные слова «00» можно расположить по порядку способами, так что общее число различных вариантов реконструкции со словами «00», построенных на основе любой допустимой последовательности из слов «10» и «01» равно $r^{n_{00}} \cdot n_{00}!$.
Аналогично разместим n_{11} слов «11» в любом из s мест между словами «01» и «10» или на конце «1» уже построенной последовательности, всего с учетом уникальности слов «11» получим $s^{n_{11}} \cdot n_{11}!$ вариантов, где

$$s = \begin{cases} m, \text{"01" } \delta \text{ "10 } \textit{поровну, соединены начиная с } \text{"01"} \\ m+1, \textit{иначе} \end{cases}. \quad (7)$$

Объединяя результаты размещения слов вида «00» и «11» по произвольной (одной) последовательности из слов «01» и «10», получим, что число возможных последовательностей, построенных вставкой слов вида «00» и слов вида «11» равно
$r^{n_{00}} \cdot n_{00}! \cdot s^{n_{11}} \cdot n_{11}!$, (8)

где и находятся по формулам (6) и (7).

Заметим, что если количество слов «01» и «10», отличается на один (слов «01» и слов «10» или слов «01» и слов «10»), $\alpha = 1$, то $r = s = m+1$, то с учетом формул (4) и (8) число реконструкций по множеству равно
$N = m! \cdot (m+\alpha)! \cdot (2-\alpha) \cdot n_{00}! \cdot n_{11}! \cdot (m+1)^{n_{00}+n_{11}+1}$. (9)

В случае если слов вида «01» и «10» по штук, $\alpha = 0$, то для последовательностей, начинающихся на «01» $r = m+1, s = m$, число реконструкций
$N = m! \cdot (m+\alpha)! \cdot (2-\alpha) \cdot n_{00}! \cdot n_{11}! \cdot (m+1)^{n_{00}} \cdot m^{n_{11}}$, (10)
а для последовательностей, начинающихся на «10» $r = m$, $s = m+1$, и число реконструкций
$N = m! \cdot (m+\alpha)! \cdot (2-\alpha) \cdot n_{00}! \cdot n_{11}! \cdot (m+1)^{n_{11}} \cdot m^{n_{00}}$, (11)
так что общее число реконструкций равно
$N = m! \cdot (m+\alpha)! \cdot (2-\alpha) \cdot n_{00}! \cdot n_{11}! \cdot \left((m+1)^{n_{00}} \cdot m^{n_{11}} + (m+1)^{n_{11}} \cdot m^{n_{00}} \right)$. (12)

Резюмируя вышеизложенное, отметим, что число реконструкций в случае, когда состоит из слов вида «01», $m+1$ слов вида «10» (или наоборот), слов вида «00» и n_{11} слов вида «11», вычисляется по формуле (9). Если мы имеем по слов вида «10» и «01», то число реконструкций находится по формуле (12) с $\alpha = 0$.

Реконструкция при разрешенных инверсиях

При разрешенных инверсиях слов мы допускаем возможности соединения как слов, содержащихся во множестве , так и их инверсий. Например, два слова «010110» и «0100» можно соединить как в прямом порядке — «010110» + «0100» = «010110100» или «0100» + «010110» = «010010110», так и в инвертированном порядке, причем инверсия возможна как для одного слова «010110» + «0010», «0010» + «010110» или «0100» + «011010», «011010» + «0100», так и для двух слов «011010» + «0010», «0010» + «011010». Отметим, что возможность инверсий существенно увеличивает число возможных реконструкций.

Особенности постановки задачи при разрешенных инверсиях связаны с наличием в симметричных и несимметричных слов. Очевидно, что симметричные слова (палиндромы) не увеличивают число реконструкций, и это могут быть только слова вида «00» и «11». Проведем вначале оценку в ситуации отсутствия палиндромов.

Будем считать, что при соединении слов вида «01» и «10» их можно инвертировать, так что из слова «01» получается «10». Для простоты считаем, что во множестве есть только слова вида «01» и нет слов вида «10», иначе переворачиваем слова «10» и получаем слова «01».

Пусть $n_{01} = 2m$, т.е. n_{01} четное, тогда половину слов «01» инвертируем и получаем по m слов вида «01» и «10». В этом случае число реконструкций находится по формуле (12).

В случае нечетного n_{01}, т.е. $n_{01} = 2m+1$ инвертируем m слов «01» и получаем число реконструкций на основе формулы (9).

Отметим, что среди этих реконструкций могут быть такие, что превращаются друг в друга с помощью переворота. Записывая полученные реконструкции в обратном порядке, получим новые реконструкции, такие же, как если перевернуть не m а $m+1$ слов «01», поэтому для подсчета общего числа возможных реконструкций нужно формулу (9) умножить на 2. Заметим, что половину слов вида «01» для инверсии в случае начального четного n_{01} можно выбрать C_{2m}^{m} способами, поэтому общее число реконструкций, полученное по формуле (12) увеличивается в C_{2m}^{m} раз, а для случая нечетного n_{01} общее число реконструкций, полученное по формуле (9) увеличивается в C_{2m+1}^{m} раз.

Оценка с учетом числа палиндромов в словах вида «00» и «11», связана с тем, что, мы допускаем не только инверсию слов вида «01» и «10», но и инверсию слов вида «00» и «11». Пусть $\hat{n}_{00}$ и $\hat{n}_{11}$ — число несимметричных слов вида «00» и «11». Такие инверсии приведут к увеличению числа реконструкций в $2^{(\hat{n}_{00}+\hat{n}_{11})}$ раз, поскольку каждое несимметричное слово вида «00» и «11» (несимметричные слова вида «00» и «11» содержат внутри несимметричную последовательность символов) можно инвертировать независимо от остальных.

Заключение

В статье рассмотрена задача получения точных оценок числа реконструкций для слов над бинарным алфавитом и подслов различной длины при сборке с наложением в один символ. Предложен подход, основанный на рассмотрении усеченных слов, состоящих из префикса и суффикса исходных слов длины один. Такой подход позволил единообразно работать с исходными словами различной длины. Указаны особенности исходного множества слов, при которых реконструкция невозможна. Рассмотрены различные случаи, допускающие полную реконструкцию (сборку), и получены формулы, дающие точное число реконструкций. Дополнительно рассмотрена ситуация реконструкции с разрешенными инверсиями слов. Дальнейшее развитие предложенного подхода авторы видят в рассмотрении сборок с наложением в несколько символов.

References

[1] Lothaire M. Combinatorics on Words. In: Encyclopedia of Mathematics and its Applications. vol. 17. Addison-Wesley, Reading, Mass; 1983. (In Eng.)

[2] Lothaire M. Algebraic Combinatorics on Words. Cambridge: Cambridge University Press; 2002. (In Eng.) DOI: https://doi.org/10.1017/CBO9781107326019

[3] Lothaire M. Applied Combinatorics on Words. In: Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press; 2005. (In Eng.)

[4] Berstel J., Karhumäki J. Combinatorics on Words – A Tutorial. *Bulletin of the European Association for Theoretical Computer Science.* 2003; 79:178-228. Available at: https://eatcs.org/images/bulletin/beatcs79.pdf (accessed 27.05.2020). (In Eng.)

[5] Kahrumäki J. Combinatorics on Words: A New Challenging Topic. Technical Report no. 645. Turku Center for Computer Science; 2004. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.537.1828&rep=rep1&type=pdf (accessed 27.05.2020). (In Eng.)

[6] Berstel J., Perrin D. The origins of combinatorics on words. *European Journal of Combinatorics*. 2007; 28(3):996-1022. (In Eng.) DOI: https://doi.org/10.1016/j.ejc.2005.07.019

[7] Blanchet-Sadri F. Algorithmic Combinatorics on Partial Words. Chapman & Hall/CRC; 2007. (In Eng.) DOI: https://doi.org/10.1201/9781420060935

[8] Lascoux A., Schützenberger M.P. A New Statistics on Words. *Annals of Discrete Mathematics*. 1980; 6:251-255. (In Eng.) DOI: https://doi.org/10.1016/S0167-5060(08)70709-X

[9] Lascoux A., Schützenberger M.-P. Sur une conjecture de H.O. Foulkes. *C. R. Acad. Sci. Paris.* 1978; 286 A(7):323-324. Available at: http://www-igm.univ-mlv.fr/~berstel/Mps/Travaux/A/1978-4FoulkesCras.pdf (accessed 27.05.2020). (In French)

[10] Reutenauer C., Schützenberger M.P. Rational word functions: Characterization and minimization. In: M. Ito (ed.) Words, Languages and Combinatorics. *Proceedings of the International Colloquium*. Kyoto, Japan, 28-31 August 1990. World Sci. Publishing, River Edge, NJ; 1992. p. 435-443. Available at: http://www-igm.univ-mlv.fr/~berstel/Mps/Travaux/A/1992-3ChristopheKyoto.pdf (accessed 27.05.2020). (In Eng.)

[11] Levenshtein V.I. Maximum Number of Words in Codes without Overlaps. *Problemy Peredachi Informatsii* = Problems of Information Transmission. 1970; 6(4):355-357. (In Eng.)

[12] Levenshtein V.I. Upper-Bound Estimates for Fixed-Weight Codes. *Problemy Peredachi Informatsii* = Problems of Information Transmission. 1971; 7(4):281-287. (In Eng.)

[13] Dress A.W.M., Erdős P.L. Reconstructing Words from Subwords in Linear Time. *Annals of Combinatorics*. 2005; 8(4):457-462. (In Eng.) DOI: https://doi.org/10.1007/s00026-004-0232-4

[14] Gamow G. Combinatorial Principles in Genetics. In: E.F. Beckenbach, G. Pólya (ed.) *Applied Combinatorial Mathematics*. New York, J. Wiley; 1964. p. 515-535. (In Eng.)

[15] Jonoska N., Saito M. (eds.) Discrete and Topological Models in Molecular Biology. Natural Computing Series. Springer, Berlin, Heidelberg; 2014. (In Eng.) DOI: https://doi.org/10.1007/978-3-642-40193-0

[16] Erdős P.L., Ligeti P., Sziklai P., Torney D.C. Subwords in Reverse-Complement Order. *Annals of Combinatorics*. 2006; 10(4):415-430. (In Eng.) DOI: https://doi.org/10.1007/s00026-006-0297-3

[17] Herbelot A., Baroni M. High-risk learning: acquiring new word vectors from tiny data. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP-2017)*. Copenhagen, Denmark, Association for Computational Linguistics; 2017. p. 304-309. (In Eng.) DOI: https://doi.org/10.18653/v1/D17-1030

[18] Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. 2017; 5:135-146. Available at: https://transacl.org/ojs/index.php/tacl/article/view/999 (accessed 27.05.2020). (In Eng.)

[19] Pinter Y., Guthrie R., and Eisenstein J. Mimicking Word Embeddings using Subword RNNs. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP-2017)*. Copenhagen, Denmark, Association for Computational Linguistics; 2017. p. 102-112. (In Eng.) DOI: https://doi.org/10.18653/v1/D17-1010

[20] Zhao J., Mudgal S., Liang Y. Generalizing Word Embeddings using Bag of Subwords. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP-2018*). Brussels, Belgium, Association for Computational Linguistics; 2018. p. 601-606. (In Eng.) DOI: https://doi.org/10.18653/v1/D18-1059

[21] Sasaki S., Suzuki J., Inui K.. Subword-based Compact Reconstruction of Word Embeddings. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL-2019)*. Minneapolis, Minnesota, Association for Computational Linguistics; 2019. p. 3498-3508. (In Eng.) DOI: https://doi.org/10.18653/v1/N19-1353

[22] Pennington J., Socher R., and Manning C. GloVe: Global Vectors for Word Represqentation. In*: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-*2014). Doha, Qatar, Association for Computational Linguistics; 2014. p. 1532-1543. (In Eng.) DOI: https://doi.org/10.3115/v1/D14-1162

[23] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781. Available at: https://arxiv.org/abs/1301.3781 (accessed 27.05.2020). (In Eng.)

[24] Smetanin Y.G., Ulyanov M.V. Reconstruction of a Word from a Finite Set of its Subwords under the unit Shift Hypothesis. I. Reconstruction without for Bidden Words. *Cybernetics and Systems Analysis*. 2014; 50(1):148-156. (In Eng.) DOI: https://doi.org/10.1007/s10559-014-9602-z

[25] Smetanin Y.G., Ulyanov M.V. Reconstruction of a Word from a Finite Set of its Subwords Under the Unit Shift Hypothesis. II. Reconstruction with Forbidden Words. *Cybernetics and Systems Analysis*. 2015; 51(1):157-164. (In Eng.) DOI: https://doi.org/10.1007/s10559-015-9708-y

[26] Smetanin Y., Ulyanov M., Shulga M. On Calculating the Entropy of 2D Words Over a Finite Alphabet. In: *2018 International Conference on Engineering Technologies and Computer Science (EnT-2018)*. Moscow; 2018. p. 82-85. (In Eng.) DOI: https://doi.org/10.1109/EnT.2018.00025

[27] Ulyanov M., Smetanin Y. Entropy Function of Finite Words. In: *2017 International Workshop on Engineering Technologies and Computer Science (EnT-2017)*. Moscow; 2017. p. 8-11. (In Eng.) DOI: https://doi.org/10.1109/EnT.2017.7

Поступила 27.05.2020; принята к публикации 14.08.2020; опубликована онлайн 30.09.2020.
Submitted 27.05.2020; revised 14.08.2020; published online 30.09.2020.

Об авторах:

Жукова Галина Николаевна, доцент департамента программной инженерии, факультет компьютерных наук, ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики» (101000, Россия, г. Москва, ул. Мясницкая, д. 20), кандидат физико-математических наук, доцент, ORCID: http://orcid.org/0000-0003-1835-7422, gzhukova@hse.ru

Сметанин Юрий Геннадиевич, главный научный сотрудник Вычислительного центра им. А.А. Дородницына РАН, ФГУ «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (119333, Россия, г. Москва,

ул. Вавилова, д. 44, корп. 2), доктор физико-математических наук, ORCID: http://orcid.org/0000-0003-0242-6972, ysmetanin@rambler.ru
Ульянов Михаил Васильевич, ведущий научный сотрудник, ФГБУН «Институт проблем управления им. В.А. Трапезникова Российской академии наук» (117997, Россия, г. Москва, ул. Профсоюзная, д. 65), профессор кафедры алгоритмических языков, факультет вычислительной математики и кибернетики, ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова» (119991, Россия, г. Москва, ГСП-1, Ленинские горы, д. 1), доктор технических наук, профессор, ORCID: http://orcid.org/0000-0002-5784-9836, muljanov@mail.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Galina N. Zhukova, Associate Professor of the School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics (20 Myasnitskaya St., Moscow 101000, Russia), Ph.D. (Phys.-Math.), Associate Professor, ORCID: http://orcid.org/0000-0003-1835-7422, gzhukova@hse.ru
Yuri G. Smetanin, Chief Researcher of the Dorodnicyn Computing Centre of RAS, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (44-2 Vavilov St., Moscow 119333, Russia), Dr. Sci. (Phys.-Math.), ORCID: http://orcid.org/0000-0003-0242-6972, ysmetanin@rambler.ru
Mikhail V. Uljanov, Leading Scientist Researcher, V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences (65 Profsoyuznaya St., Moscow 117997, Russia), Professor of the Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1, Leninskie gory, Moscow 119991, Russia), Dr.Sci. (Engineering), Professor, ORCID: http://orcid.org/0000-0002-5784-9836, muljanov@mail.ru

All authors have read and approved the final manuscript.