Теория кодирования

МФТИ, осень 2013

Александр Дайняк

www.dainiak.com

Теория кодирования

Теория кодирования изучает модели хранения и передачи «дискретной» информации и предлагает способы оптимального её кодирования.

01101010101001000100001010 1

Алфавитное кодирование

Как хранить в компьютере тексты на естественном языке?

Нужно их кодировать!

Простейший подход:

- каждой букве языка, а также знакам препинания сопоставим по двоичному слову,
- и тогда текст закодируем, записав друг за другом коды отдельных букв.

Алфавитное кодирование

Математическая модель:

Даны алфавиты

$$\mathbb{A} = \{a_1, \dots, a_n\} \quad \text{и} \quad \mathbb{B} = \{b_1, \dots, b_q\}$$

Алфавит $\mathbb{A}$ — *кодируемый*, «естественный»;
алфавит $\mathbb{B}$ — *кодовый* (например, $\mathbb{B} = \{0,1\}$).

Алфавитное кодирование — это отображение

$$\phi: \mathbb{A}^* \to \mathbb{B}^*$$

такое, что для любых $a_{i_1}, \dots, a_{i_r}$ выполнено

$$\phi(a_{i_1} \dots a_{i_r}) = \phi(a_{i_1}) \dots \phi(a_{i_r})$$

Алфавитное кодирование

Достаточно определить ϕ на отдельных символах алфавита A:

$$\begin{gathered}\phi(a_1) = B_1 \\ \vdots \\ \phi(a_n) = B_n\end{gathered}$$

Слова $B_1, \ldots, B_n$ называются *кодовыми*, совокупность $\{B_1, \ldots, B_n\}$ называется *кодом*.

Везде далее считаем, что все B_i различны, иначе кодирование не *однозначное*.

Но этого в общем случае недостаточно…

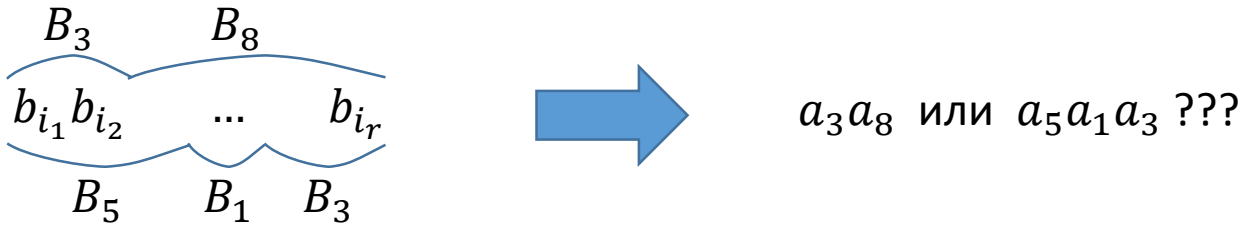
Однозначность алфавитного кодирования

Кодирование ϕ *однозначное*, если

$$\phi(w') \neq \phi(w'') \text{ при } w' \neq \mathrm{w}''$$

Однозначность никак не зависит от алфавита $\mathbb{A}$, а целиком определяется набором $\{B_1, \dots, B_n\}$.

Кодирование однозначное т. и т.т., когда никакое слово $b_{i_1} b_{i_2} \dots b_{i_r}$ нельзя двумя разными способами разбить на кодовые слова:



Достаточные условия однозначности

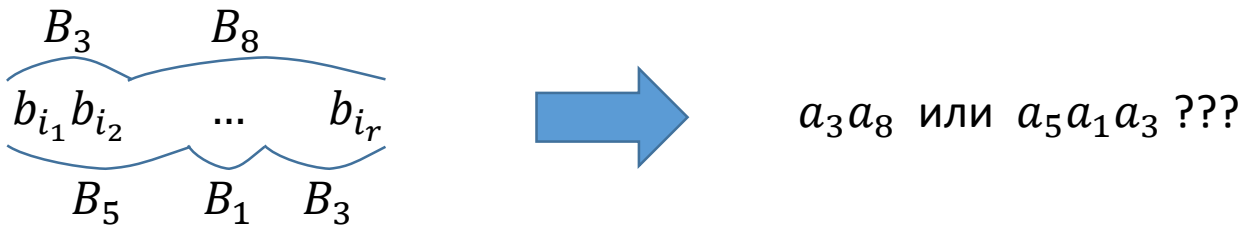
- Равномерность: $|B_1| = |B_2| = \cdots = |B_n|$
- Свойство префикса:

$$\nexists i, j \left(i \neq j \text{ и } B_i = B_j w, \text{где } w \in \mathbb{B}^* \right)$$

- Свойство суффикса:

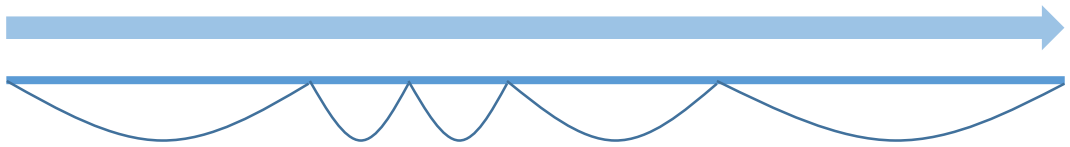
$$\nexists i, j \left(i \neq j \text{ и } B_i = w B_j, \text{где } w \in \mathbb{B}^* \right)$$

При этом такой ситуации не возникнет:



Префиксные коды — «мгновенные»

Префиксные коды называют ещё *мгновенными*, так как закодированные с их помощью сообщения можно декодировать по мере приёма, без задержек:



Нужен критерий однозначности!

Равномерность, префиксность и суффиксность не являются необходимыми условиями для однозначности. Пример:

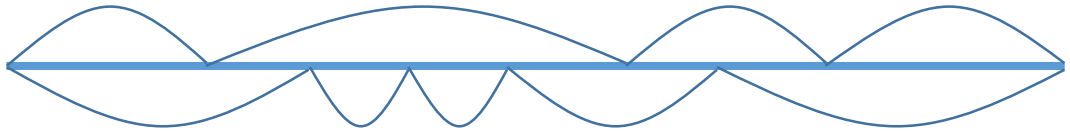
$$\mathbb{A} = \{a_1, a_2\}, \mathbb{B} = \{0,1\}$$
$$\phi(a_1) = 0$$
$$\phi(a_2) = 010$$

Полезно было бы получить *критерий* однозначности кода.

Вывод критерия однозначности

Код неоднозначен, если найдётся слово $B \in \mathbb{B}^*$, которое не менее чем двумя разными способами можно разбить на кодовые слова.

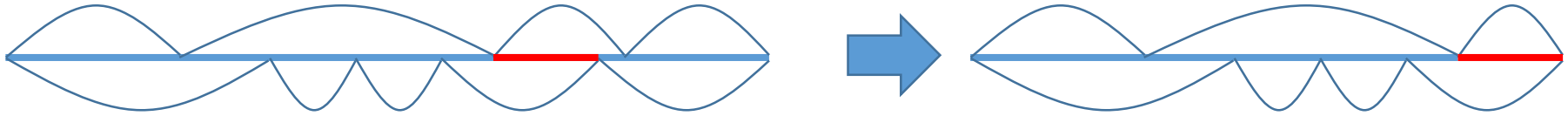
Рассмотрим самое короткое такое «неоднозначное» B и два его различных разбиения на кодовые слова:



Вывод критерия однозначности

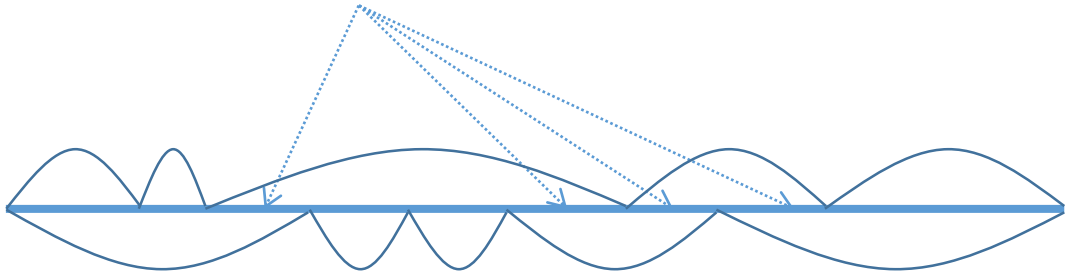
Заметим, что точки «верхнего» и «нижнего» разбиений, кроме крайних, все различны, иначе слово B можно было бы укоротить:

Также, среди отрезков B, концы которых принадлежат разным разбиениям, нет кодовых слов, иначе B также можно было бы укоротить:



Вывод критерия однозначности

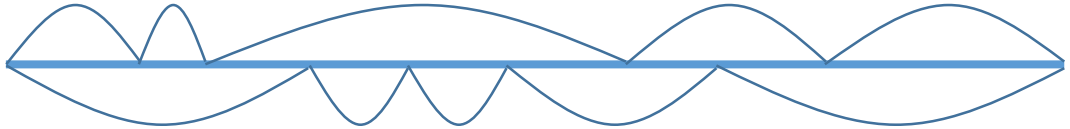
Минимальные отрезки слова B, концы которых принадлежат разным разбиениям, назовём *промежуточными*.



Первый пром. отрезок получается, если из начала некоторого кодового слова «отнять» некоторую последовательность кодовых слов.

Любой из остальных отрезков получается, если из некоторого кодового слова отнять предыдущий отрезок и последовательность (возможно, пустую) кодовых слов.

Вывод критерия однозначности



Последний пром. отрезок таков, что если его отнять из начала некоторого кодового слова, получится последовательность кодовых слов.

Обозначим через $w_1, \ldots, w_k$ все промежуточные отрезки.

Через β будем обозначать последовательность (возможно, пустую) кодовых слов. Имеем:

$$\exists i, \beta(B_i = \beta w_1)$$
$$\exists i, \beta(B_i = w_1 \beta w_2)$$
$$\vdots$$
$$\exists i, j(B_i = w_{k-1} \beta w_k)$$
$$\exists i, j(B_i = w_k \beta)$$

Вывод критерия однозначности

Наоборот, пусть нашлись непустые слова $w_1, \dots, w_k \in \mathbb{B}^*$, кодовые слова $B_{i_1}, \dots, B_{i_{k+1}}$ и последовательности кодовых слов $\beta_1, \dots, \beta_{k+1}$, такие, что выполнены соотношения

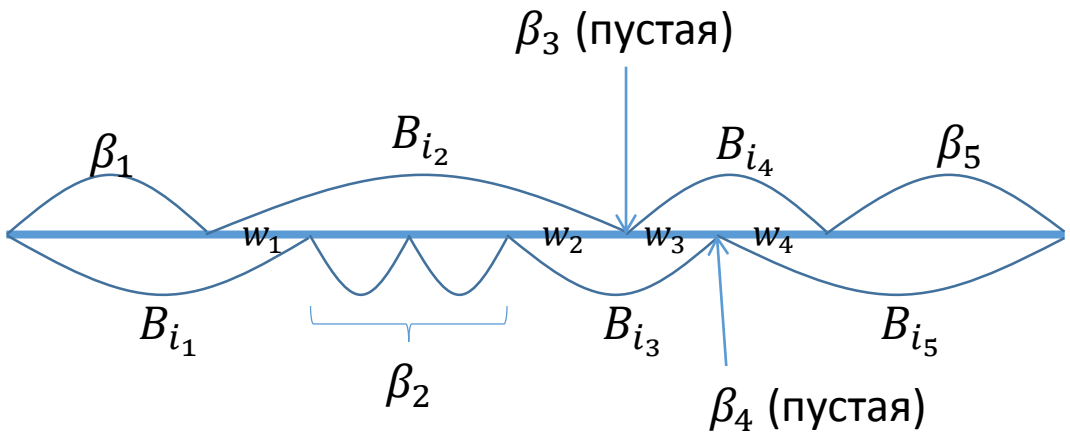
$$B_{i_1} = \beta_1 w_1$$
$$B_{i_2} = w_1 \beta_2 w_2$$
$$\vdots$$
$$B_{i_{k+1}} = w_k \beta_{k+1}$$

Тогда слово $\beta_1 w_1 \beta_2 w_2 \dots w_{k-1} \beta_k w_k \beta_{k+1}$ можно разбить на кодовые слова двумя способами.

Вывод критерия однозначности

$$B_{i_1} = \varepsilon\beta_1 w_1$$
$$B_{i_2} = w_1\beta_2 w_2$$
$$\vdots$$
$$B_{i_{k+1}} = w_k\beta_{k+1}\varepsilon$$

Пример:



Критерий однозначности алфавитного кодирования

Код $C = \{B_1, \dots, B_n\}$ *не* однозначный т. и т.т., когда найдутся непустые слова $w_1, \dots, w_k \in \mathbb{B}^* \setminus C$, кодовые слова $B_{i_1}, \dots, B_{i_{k+1}}$ и последовательности кодовых слов $\beta_1, \dots, \beta_{k+1}$, такие, что $k \geq 1$ и выполнены соотношения

$$
\begin{aligned}
B_{i_1} &= \beta_1 w_1 \\
B_{i_2} &= w_1 \beta_2 w_2 \\
&\vdots \\
B_{i_{k+1}} &= w_k \beta_{k+1}
\end{aligned}
$$

(или $k = 0$ и $B_{i_1} = \beta_1$, где β_1 составлено не менее чем из двух кодовых слов).

Ещё одна формулировка критерия однозначности

Через ε будем обозначать пустое слово.

Пусть $C = \{B_1, \ldots, B_n\}$ — код, который нужно проверить на однозначность.

Построим орграф $G_C = (V, E)$, где
$V = \{\varepsilon$, а также все слова из $\mathbb{B}^* \setminus C$, являющиеся началами и концами кодовых слов$\}$,

$$E = \left\{(\alpha', \alpha'') \mid \exists \beta \in C^* \left(\alpha' \beta \alpha'' \in C \text{ и при этом } \begin{bmatrix} \beta \neq \varepsilon \\ \alpha' \neq \varepsilon \text{ и } \alpha'' \neq \varepsilon \end{bmatrix} \right) \right\}.$$

Код C однозначный т. и т.т., когда в орграфе G_C нет орцикла, проходящего через вершину ε.

Оценка длины неоднозначно декодируемого слова

$V = \{\varepsilon$, а также все слова из $\mathbb{B}^* \setminus C$, являющиеся началами и концами кодовых слов$\}$,

$$E = \left\{(\alpha', \alpha'') \mid \exists \beta \in C^* (\alpha' \beta \alpha'' \in C) \text{и при этом} \begin{bmatrix} \beta \neq \varepsilon \\ \alpha' \neq \varepsilon \text{ и } \alpha'' \neq \varepsilon \end{bmatrix}\right\}.$$

Имеем

$$|V| \leq 1 + \sum_{B \in C} (|B| - 1) \leq |C| \cdot \max_{B \in C} |B|$$

Получим отсюда оценку длины минимального неоднозначно декодируемого слова…

Оценка длины неоднозначно декодируемого слова

Если в G_C есть цикл через ε, то есть и цикл, число вершин в котором не больше, чем

$$|C| \cdot \max_{B \in C} |B|$$

Рассмотрим соответствующее этому циклу неоднозначно декодируемое слово

$$W_{\text{неодн.}} = \beta_1 w_1 \beta_2 w_2 \dots w_{k-1} \beta_k w_k \beta_{k+1}$$

Каждая пара $\beta_i w_i$ умещается в некотором кодовом слове, поэтому

$$\left|W_{\text{неодн.}}\right| \leq (k+1) \cdot \max_{B \in C} |B| \leq |C| \cdot \left(\max_{B \in C} |B|\right)^2$$

Оценка длины неоднозначно декодируемого слова

Из предыдущих рассуждений вытекает оценка на длину неоднозначно декодируемого слова:

Теорема. (А.А. Марков)
Если C — неоднозначный код, длина слов которого не превосходит l, то найдётся слово длины не более $|C| \cdot l^2$, декодируемое неоднозначно.

Коды с минимальной избыточностью

Обычно, кодируемые символы $a_1, \dots, a_n$ встречаются в кодируемых сообщениях не одинаково часто, а с разными частотами.

Например, в английском языке буква *e* встречается примерно в 180 раз чаще, чем *z*.

Естественно при построении кодирования ϕ кодировать более частые буквы более короткими словами.

Поставим задачу математически…

Коды с минимальной избыточностью

Пусть в кодируемых сообщениях символы $a_1, \dots, a_n$ встречаются с частотами $p_1, \dots, p_n$ соответственно. Считаем $\sum p_i = 1$ и $\forall i\ p_i > 0$.

Пусть символ a_i кодируется словом B_i.

Рассмотрим сообщение $A \in \mathbb{A}^*$.

Каждый из символов a_i встретится в $|A|$ примерно $|A| \cdot p_i$ раз. Отсюда

$$|\phi(A)| \approx \sum_i |A| \cdot p_i \cdot |B_i| = |A| \cdot \sum_i p_i \cdot |B_i|$$

Коды с минимальной избыточностью

То есть, «среднестатистическое» сообщение A при кодировании «разбухает» примерно в $\sum_i p_i|B_i|$ раз.

Величина $\sum_i p_i|B_i|$ называется *коэффициентом избыточности кода*.

Задача построения кода с минимальной избыточностью:
По заданным $p_1, \dots, p_n$ построить (однозначно декодируемый!) код $B_1, \dots, B_n \in \mathbb{B}^*$, для которого коэффициент избыточности минимален.

Такой код называется *кодом с минимальной избыточностью для набора частот* $p_1, \dots, p_n$.

Неравенство Крафта—Макмиллана

Все слова кода не получится взять слишком короткими, иначе код не будет однозначным. Количественно это выражает

Теорема. (L.G. Kraft, B. McMillan)

Пусть $l_1, \dots, l_n$ — длины слов однозначного кода в алфавите $\mathbb{B}$, где $|\mathbb{B}| = q$.

Тогда выполнено неравенство

$$\sum_{i=1}^{n} q^{-l_i} \leq 1$$

Неравенство Крафта—Макмиллана

Доказательство теоремы:

Пусть $B_1, \dots, B_n$ — однозначный код в q-значном алфавите, и пусть $|B_i| = l_i$.

Пусть $t \in \mathbb{N}$. Обозначим $L \coloneqq t \cdot \max\limits_i l_i$.

Рассмотрим выражение

$$\left(\sum_{i=1}^{n} q^{-l_i}\right)^t = \sum_{1 \le i_1, \dots, i_t \le n} q^{-(l_{i_1} + \dots + l_{i_t})} = \sum_{l=1}^{L} s_l q^{-l},$$

где s_l — количество наборов $(i_1, \dots, i_t)$, таких, что $l_{i_1} + \dots + l_{i_t} = l$.

Неравенство Крафта—Макмиллана

$s_l = |S_l|$, где $S_l = \{(i_1, \dots, i_t) \mid l_{i_1} + \dots + l_{i_t} = l\}$

Каждому набору $(i_1, \dots, i_t) \in S_l$ поставим в соответствие слово $B_{i_1} \dots B_{i_t} \in \mathbb{B}^*$.

Тогда *разным наборам из S_l соответствуют разные слова* (т.к. код однозначный).

Отсюда $s_l \leq q^l$ и следовательно

$$\left(\sum_{i=1}^{n} q^{-l_i}\right)^t = \sum_{l=1}^{L} s_l q^{-l} \leq \sum_{l=1}^{L} 1 = L.$$

Неравенство Крафта—Макмиллана

Получили, что для любого $t \in \mathbb{N}$ выполнено

$$\sum_{i=1}^{n} q^{-l_i} \leq \left(t \cdot \max_i l_i\right)^{1/t}$$

Устремляя t к бесконечности, получаем

$$\sum_{i=1}^{n} q^{-l_i} \leq 1$$

Существование префиксных кодов

Докажем обратное утверждение:

Теорема.
Пусть натуральные числа $l_1, \ldots, l_n$ и q таковы, что

$$\sum_{i=1}^{n} q^{-l_i} \leq 1$$

Тогда существует *префиксный* код $B_1, \ldots, B_n$
в q-значном алфавите, такой, что $|B_i| = l_i$.

Существование префиксных кодов

Пусть натуральные числа $l_1, \dots, l_n$ и q таковы, что

$$\sum_{i=1}^{n} q^{-l_i} \leq 1$$

Будем считать, что среди $l_1, \dots, l_n$ всего m различных, и при этом $l_1 < \dots < l_m$. Для каждого $j \in [1, m]$ положим

$$n_j := \left|\left\{i \in [1, n] \mid l_i = l_j\right\}\right|$$

Тогда из условия теоремы следует неравенство

$$\sum_{j=1}^{m} n_j q^{-l_j} \leq 1$$

Существование префиксных кодов

Имеем неравенство $\sum_{j=1}^{m} n_j q^{-l_j} \leq 1$.

Отсюда $\sum_{j=1}^{k} n_j q^{-l_j} \leq 1$ для любого $k \in [1, m]$.

Домножив обе части на q^{l_k}, получим

$$q^{l_k} \geq \sum_{j=1}^{k} n_j q^{l_k - l_j} = n_k + \sum_{j=1}^{k-1} n_j q^{l_k - l_j}$$

Следовательно, для любого $k \in [1, m]$ имеем

$$n_k \leq q^{l_k} - \sum_{j=1}^{k-1} n_j q^{l_k - l_j}.$$

Существование префиксных кодов

Будем строить префиксный код, сначала выбирая n_1 слов длины l_1, затем n_2 слов длины l_2, и т.д.

Пусть уже набраны все кодовые слова с длинами $l_1, \dots, l_{k-1}$.

Слов длины l_k, для которых выбранные кодовые слова являются префиксами, не более $n_1 q^{l_k - l_1} + \cdots + n_{k-1} q^{l_k - l_{k-1}}$, то есть «пригодных для выбора» слов длины l_k не меньше, чем

$$q^{l_k} - \left(n_1 q^{l_k - l_1} + \cdots + n_{k-1} q^{l_k - l_{k-1}}\right)$$

Существование префиксных кодов

Пусть уже набраны все кодовые слова с длинами $l_1, \dots, l_{k-1}$.
«Пригодных для выбора» слов длины l_k не меньше, чем

$$q^{l_k} - \left(n_1 q^{l_k - l_1} + \cdots + n_{k-1} q^{l_k - l_{k-1}}\right).$$

Из условия теоремы мы ранее вывели, что

$$n_k \leq q^{l_k} - \left(n_1 q^{l_k - l_1} + \cdots + n_{k-1} q^{l_k - l_{k-1}}\right),$$

то есть мы сможем выбрать n_k слов длины l_k, так, чтобы никакие из ранее выбранных слов не были их префиксами.

По индукции получаем утверждение теоремы.

Универсальность префиксных кодов

Следствие из двух доказанных теорем.
Для любого однозначного кода существует префиксный код *в том же алфавите и с теми же длинами кодовых слов*.

Значит, *к.м.и. можно искать только среди префиксных кодов*.

Свойства оптимальных кодов

Вернёмся к задаче построения к.м.и.

Лемма.
Если $B_1, \dots, B_n$ — к.м.и. для набора частот $p_1, \dots, p_n$, то

$$\forall i, j \left(p_i > p_j \Rightarrow |B_i| \leq |B_j|\right)$$

Доказательство:
В противном случае, поменяв B_i и B_j местами, получили бы код с коэффициентом избыточности

$$\sum_{i=1}^{n} p_i|B_i| - \left(p_i - p_j\right)\left(|B_i| - \left|B_j\right|\right) < \sum_{i=1}^{n} p_i|B_i|$$

Теорема «о редукции»

Теорема «о редукции». (D.A. Huffman)
Пусть $p_1 \geq \cdots \geq p_{n-1} \geq p_n$ и $p := p_{n-1} + p_n$.

Если $B_1, \ldots, B_{n-2}, B \in \{0,1\}^*$ — префиксный к.м.и. для частот $p_1, \ldots, p_{n-2}, p$,

то $B_1, \ldots, B_{n-2}, B0, B1$ — префиксный к.м.и. для частот $p_1, \ldots, p_n$.

Доказательство теоремы о редукции

Пусть к.и. кода $B_1, \dots, B_{n-2}, B$ для частот $p_1, \dots, p_{n-2}, p$ равен k.

К.и. кода $B_1, \dots, B_{n-2}, B0, B1$ для частот $p_1, \dots, p_n$ равен

$$\sum_{i=1}^{n-2} p_i|B_i| + (p_{n-1} + p_n)(|B| + 1) = \sum_{i=1}^{n-2} p_i|B_i| + p|B| + p = k + p$$

Допустим, что нашёлся код $B'_1, \dots, B'_n$, к.и. которого для частот $p_1, \dots, p_n$ равен $k' < k + p$.

Доказательство теоремы о редукции

Допустим, что нашёлся код $B'_1, \dots, B'_n$, к.и. которого для набора частот $p_1, \dots, p_n$ равен $k' < k + p$.

Б.о.о. будем считать код $\{B'_i\}_{i=1}^n$ префиксным к.м.и. для набора $p_1, \dots, p_n$.

Т.к. $p_1 \geq \dots \geq p_n$, то $|B'_1| \leq \dots \leq |B'_n|$.

Пусть $B'_n = B'0$, где B' — некоторое слово.

Заметим, что

- $B' \notin \{B'_i\}_{i=1}^n$
- B' является префиксом одного из слов $B'_1, \dots, B'_{n-1}$.

Б.о.о. будем считать, что $B'_{n-1} = B'1$. Тогда код $B'_1, \dots, B'_{n-2}, B'$ префиксный.

Доказательство теоремы о редукции

К.и. кода $B'_1, \dots, B'_{n-2}, B'$ для набора частот $p_1, \dots, p_{n-2}, p$ равен

$$(p_{n-1} + p_n)|B'| + \sum_{i=1}^{n-2} p_i|B'_i| = \sum_{i=1}^{n} p_i|B'_i| - (p_{n-1} + p_n) = k' - p < k$$

— противоречие с тем, что код $B_1, \dots, B_{n-2}, B$ является к.м.и. для частот $p_1, \dots, p_{n-2}, p$.

Коды, исправляющие ошибки

Основные требования к кодам:

- **Однозначность** — обязательное требование. Есть критерий, алгоритм проверки.
- **Минимальная избыточность**. Есть алгоритм построения для произвольного заданного набора частот.
- **Устойчивость к ошибкам**. Возможность расшифровать закодированное сообщение даже при возникновении ошибок при его передаче.

Коды, исправляющие ошибки

Естественный язык весьма устойчив к ошибкам:

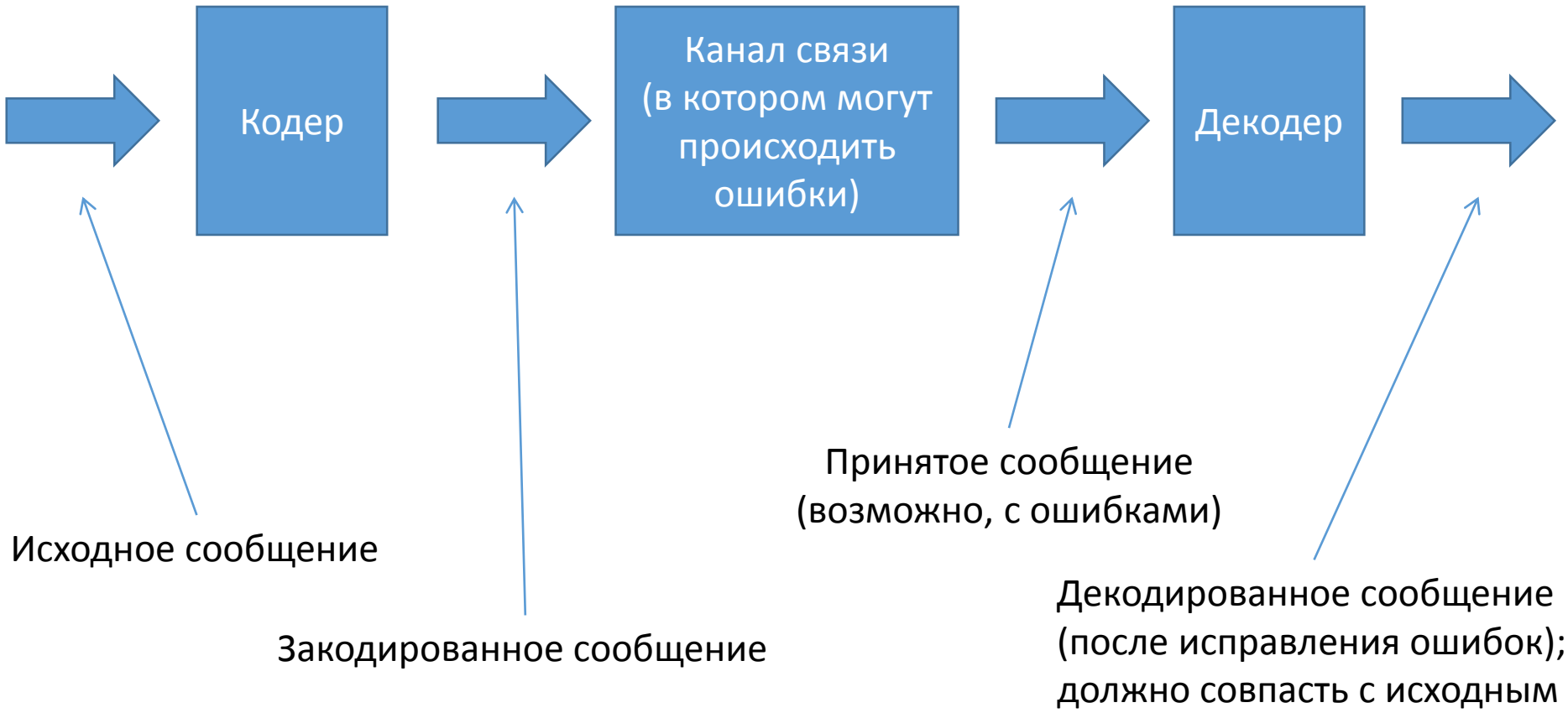
«Веть Ву мжте прчтттть эт ткст п поняц го!»

Причины:

- избыточность: гласные и т.д.
- разреженность: «вблизи» слов обычно нет других слов — если есть, то ошибки исправлять тяжело:
  чемодан зарыт vs. *чемодан закрыт*

Коды, исправляющие ошибки

Основная модель канала связи:



Что было и что будет

На лекции мы рассмотрели:

- Алфавитное кодирование
- Префиксные коды, однозначность
- Коды с минимальной избыточностью