

2025/9/4 第204回アルゴリズム研究発表会

LZ-Start-End圧縮

柴田 紘希 (九州大学)

中島 祐人 (九州大学)

山口 勇太郎 (大阪大学)

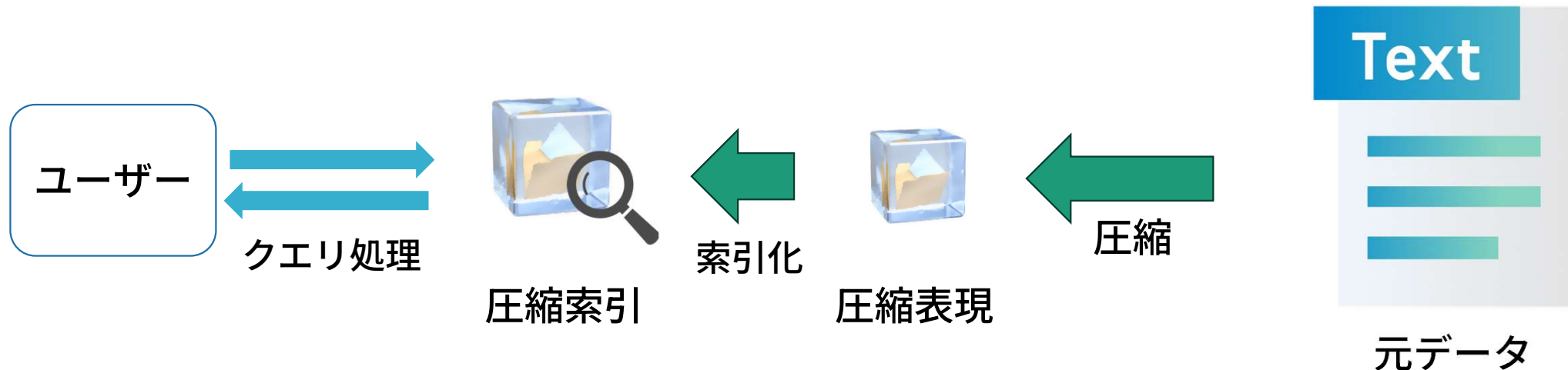
稲永 俊介 (九州大学)

データ圧縮と圧縮索引

■ データ圧縮: データを小さな圧縮表現に変換する手法

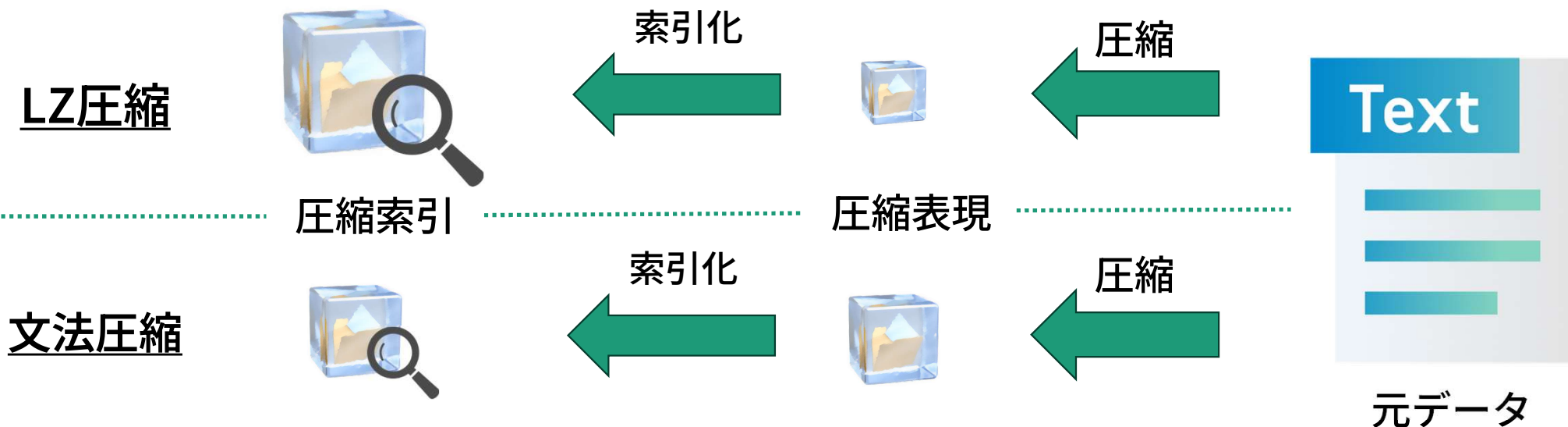
- 例: ランレングス圧縮・LZ圧縮・文法圧縮・etc...

■ 圧縮索引: 圧縮表現の構造を利用した省領域な索引構造



圧縮性能と圧縮索引化のしやすさの関係

- 圧縮性能・圧縮索引化のしやすさはトレードオフ
- 小さい圧縮表現であるほど、サイズを維持した圧縮索引化は難しい



LZ分解・文法圧縮の関係性

■ 圧縮性能: LZ圧縮の方が良い

- LZ圧縮サイズ ≤ 文法圧縮サイズ が任意の文字列で成り立つ
- LZ圧縮サイズ ≪ 文法圧縮サイズ となる文字列も存在

■ 圧縮索引化のしやすさ: 文法圧縮の方が良い

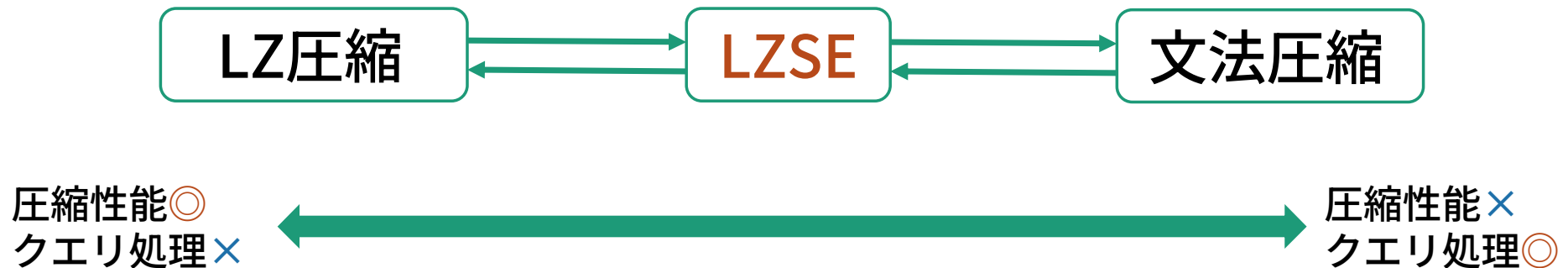
- LZ圧縮: サイズを維持した圧縮索引化は未解決
- 文法圧縮: データアクセス・文字列検索などがサイズそのまま可能



本研究の成果: LZ-Start-End (LZSE) 圧縮の提案

LZSE: 圧縮性能とクエリ処理能力を両立した制約付きLZ圧縮

- 圧縮性能: LZ圧縮～文法圧縮の間に位置
- 圧縮索引: 圧縮前データへの高速アクセスが線形サイズの圧縮索引で可能



研究の成果・発表の内容

1. LZSEと他の圧縮手法の関係性の解析

- LZSE ⇔ 文法圧縮の変換手法
- LZSEと文法圧縮の圧縮力の差の理論解析

本発表ではこの部分だけ紹介

2. 分解サイズの線形領域での圧縮索引を設計

3. 貪欲LZSE分解の線形時間アルゴリズムを提案

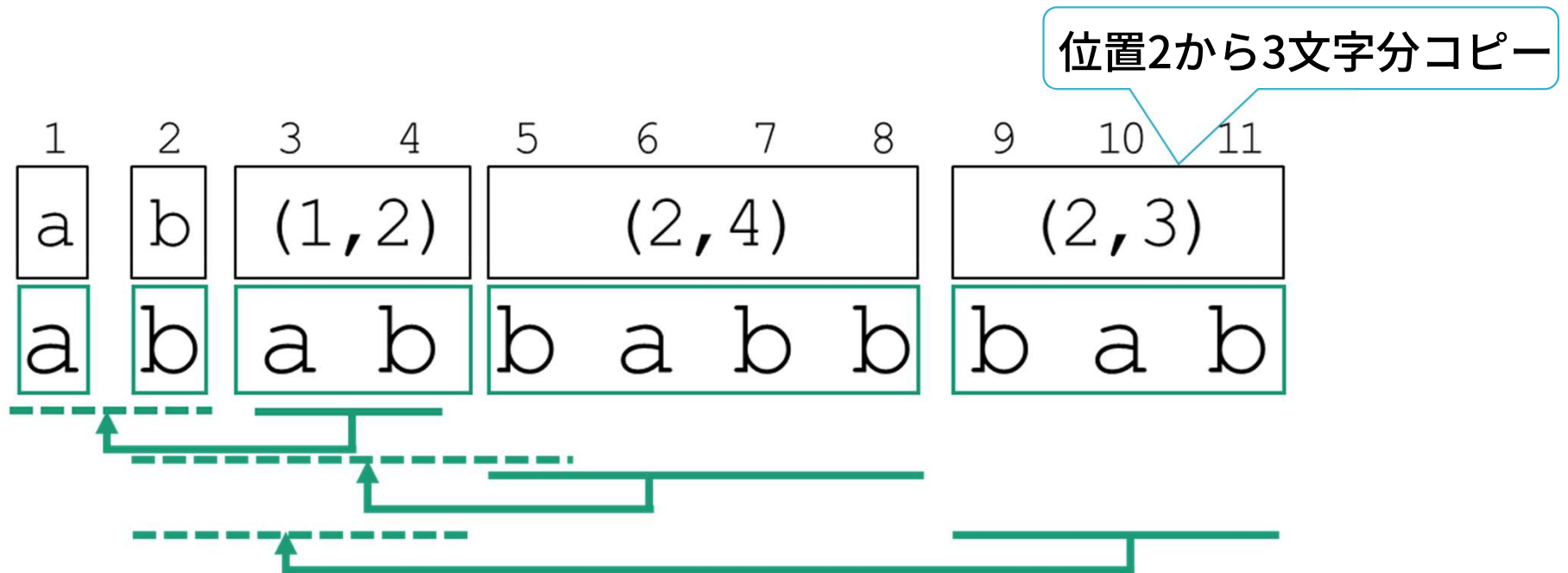
4. 貪欲LZSEの非最適性と最適LZSEとの差の下界を示す

etc...

Lempel-Ziv (LZ) 圧縮 [Ziv and Lempel, 1977]

データを以下のいずれかを表すfactorの列（LZ分解）で表現:

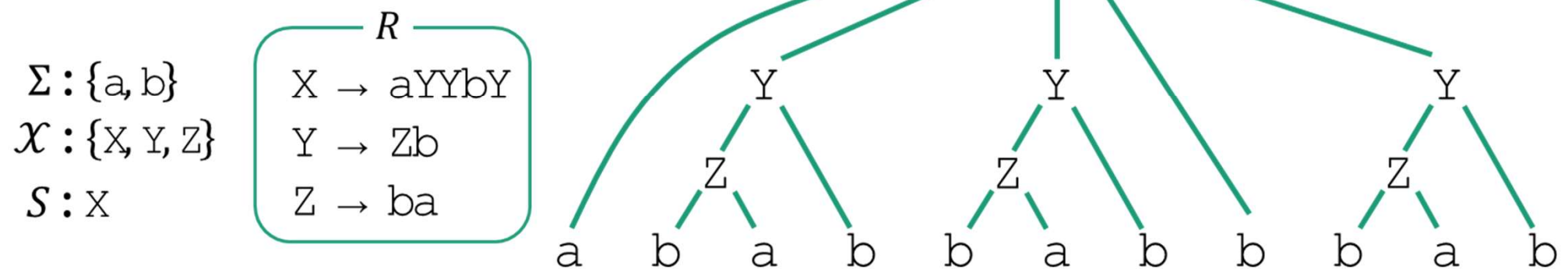
1. 1文字の保存
2. 左側に現れる同じ部分文字列へのコピー



文法圧縮 [Kieffer and Yang, 2000]

データを以下の要素からなる文法で表す圧縮手法

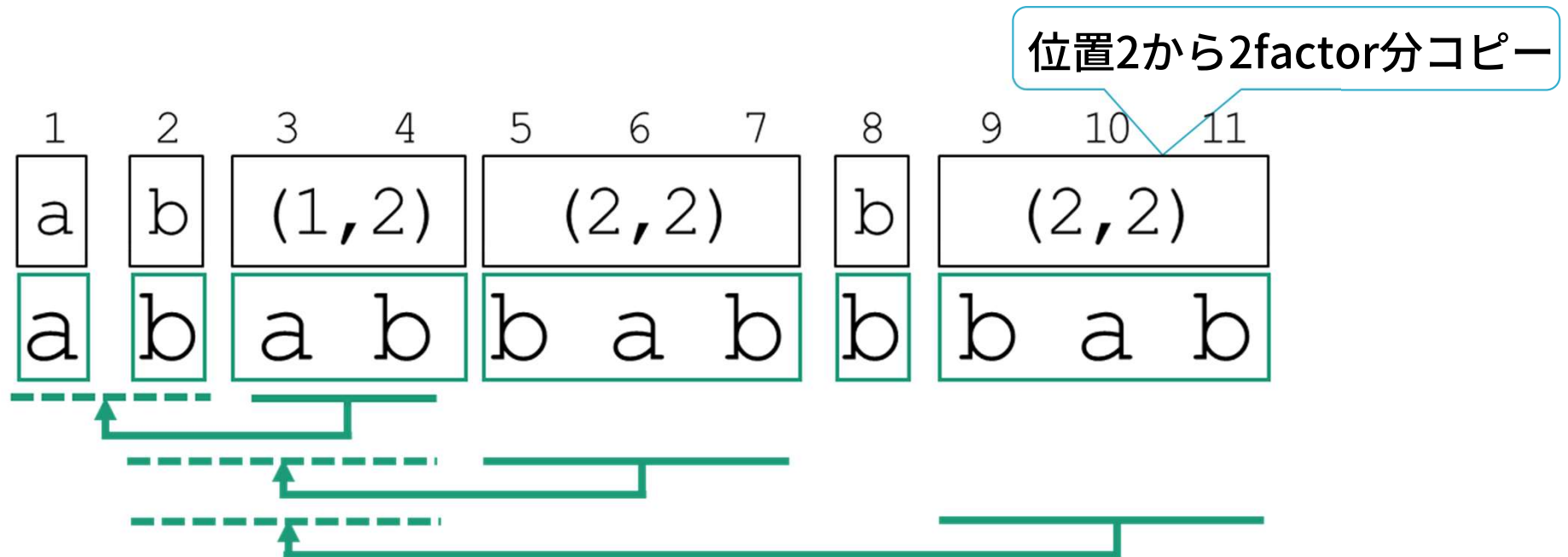
1. 終端記号（文字）の集合 Σ と、非終端記号の集合 $\mathcal{X}$
2. 非終端記号 $\rightarrow$ 記号 への変換規則の集合 R
3. 開始記号 $S \in \mathcal{X}$



新たな圧縮手法: LZ-Start-End (LZSE) 圧縮

データを以下のいずれかを表すfactorの列（**LZSE分解**）で表現:

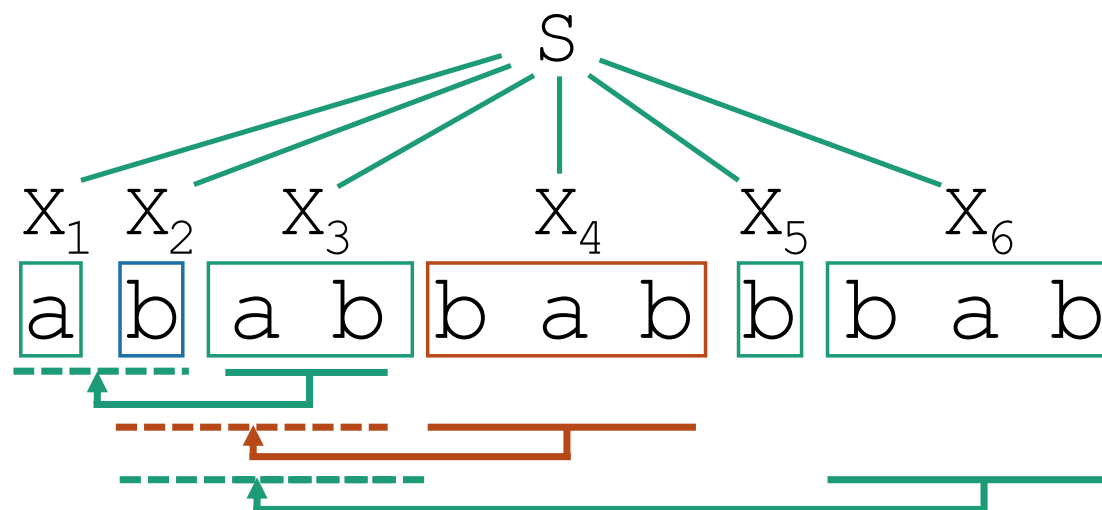
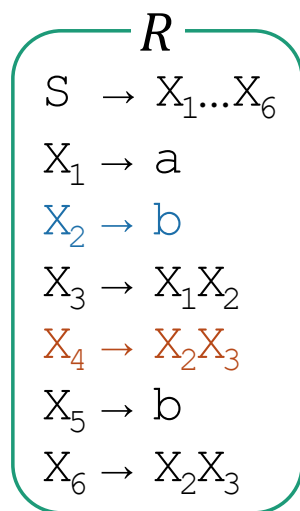
- 1文字の保存
- 左側に現れる連続するfactorへのコピー



文法への変換と圧縮索引化

各factorを1つの非終端記号に対応付けて、LZSE分解から文法を構成できる

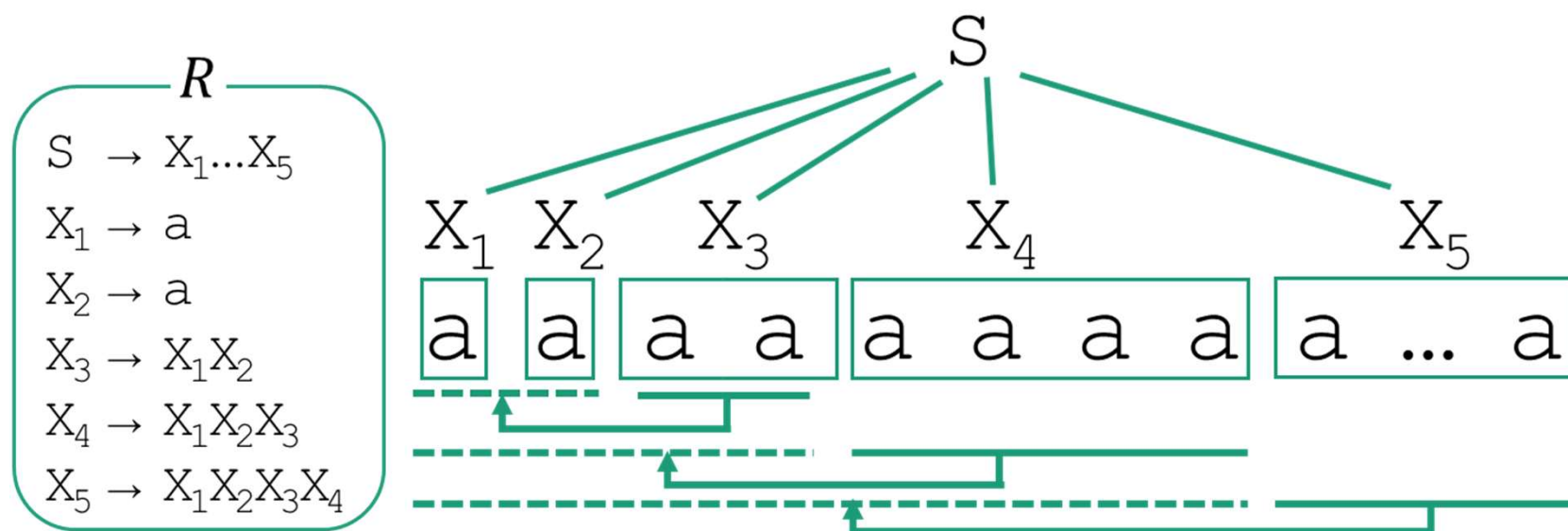
→ 文法に似た構造を持つことから、文法への手法の拡張で圧縮索引が作れる!
(詳細は略)



文法への変換によるサイズの変化

前ページの方法で構成した文法はLZSEよりサイズが大きくなるかも

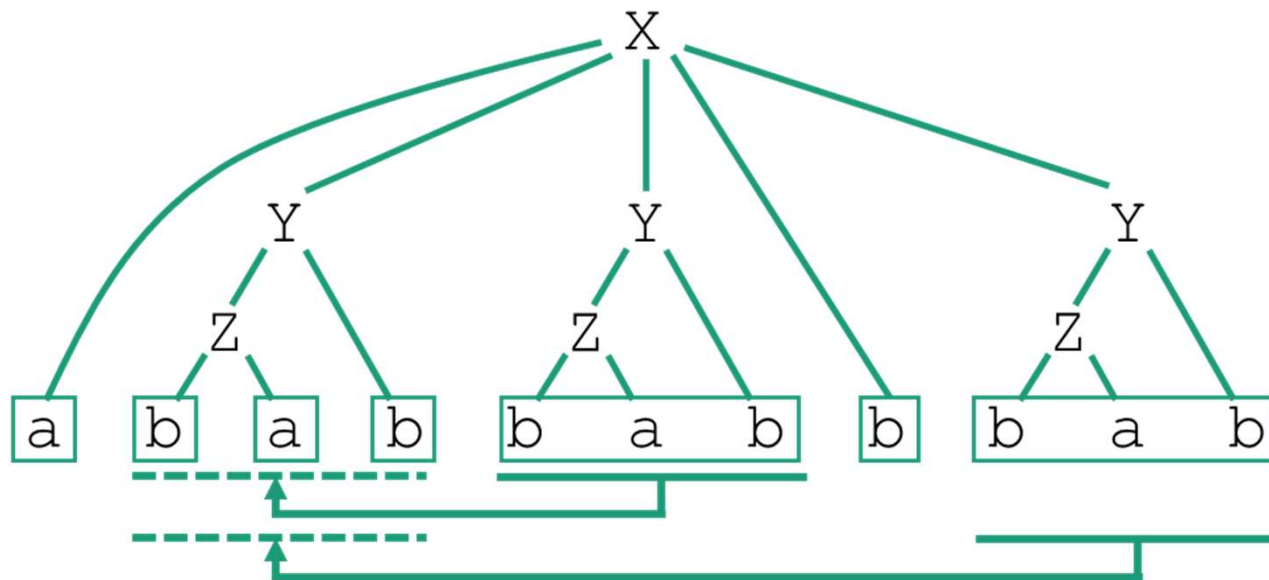
■ 下のような例が最悪ケースで、サイズが二乗オーダーで増えてしまう



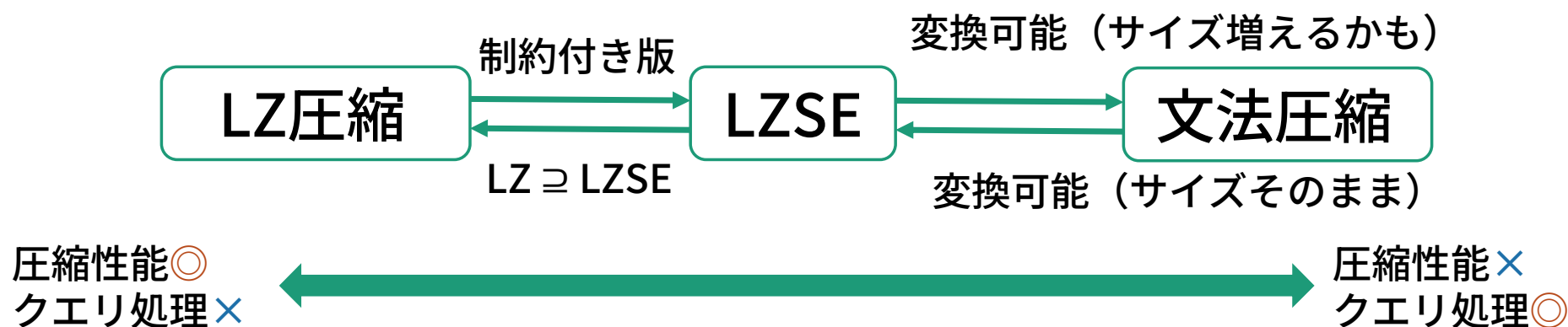
文法 → LZSE分解

Grammar Decomposition [Rytter, 2003]: 文法→LZ分解の変換手法

- 各非終端記号の非最左出現を、その記号の最左出現へのコピーとして表す
- 変換で得られたLZ分解のサイズは、元の文法圧縮のサイズ以下
- 実は、この手法で得られるLZ分解がLZSE分解になっている！



LZ・LZSE・文法圧縮の関係性



疑問: LZSEと他の圧縮表現の圧縮能力は等価なのか？

圧縮表現の最小サイズは
オーダーレベルで常に同じ？

- LZ/LZSEには、データ長 n に対しfactor数が $\Theta(\log n)$ 倍変化する例が存在
- LZSE/最小文法の間にも、同様に差が生まれる文字列を構成できないか？

文法圧縮・LZSE圧縮の圧縮性能の差

定理

十分大きな整数 m について、以下の条件を満たす長さ $\Theta(m^2)$ の文字列 T が存在する:

- T の貪欲LZSE分解のサイズが $\Theta(m)$
- T の最小文法のサイズが $\Omega(m\alpha(m))$

この定理より、LZSEが文法圧縮より（オーダーレベルで）**真に強い圧縮性能を持っている**といえる！

- 任意の文字列について、最小LZSE分解サイズ $\leq$ 最小文法圧縮サイズ
- 特定の文字列について、最小LZSE分解サイズ $\ll$ 最小文法圧縮サイズ

$\alpha(x)$: アッカーマン関数の逆関数

最小文法のサイズ下界 (1)

最小文法サイズの上界を与えるため、以下の問題を考える:

オフライン半群区間積問題

- **入力:** 要素列 $x_1, \dots, x_m \in X^m$, 区間の組 $[l_1, r_1], \dots, [l_m, r_m]$
- **出力:** $q_i = \bigotimes_{j=l_i}^{r_i} x_j$ ($1 \leq i \leq m$)

($\bigotimes$ は集合 X 上の半群演算)

既知の計算回数下界: 求解に $\Omega(m\alpha(m))$ 回の半群演算が必要な入力が存在

[Chazelle and Rosenberg, 1991]

最小文法のサイズ下界 (2)

問題の入力 $x_1, \dots, x_m \in X^m, [l_1, r_1], \dots, [l_m, r_m]$ から、以下のように文字列 T を構成する:

$$T = x_1 \cdots x_m \$_1 Q_1 \$_2 Q_2 \$_3 \cdots \$_m Q_m \$_{m+1}$$

$$Q_i = x_{l_i} \cdots x_{r_i} \quad (1 \leq i \leq m)$$

$\$_i$: 区切り文字

例: 要素列 (1,2,3,4) クエリ $[2, 3], [1, 3], [2, 4], [3, 4]$ の場合

$$T = 1234 \$_1 23 \$_2 123 \$_3 234 \$_4 34 \$_5$$

最小文法のサイズ下界 (3)

問題の入力 $x_1, \dots, x_m \in X^m, [l_1, r_1], \dots, [l_m, r_m]$ から、以下のように文字列 T を構成する:

$$T = x_1 \cdots x_m \$_1 Q_1 \$_2 Q_2 \$_3 \cdots \$_m Q_m \$_{m+1}$$

$$Q_i = x_{l_i} \cdots x_{r_i} \quad (1 \leq i \leq m)$$

$\$_i$: 区切り文字

T を導出するサイズ g の文法が存在するとき、 $O(g)$ 回の半群演算で問題が解けることを示す

→ 計算回数下界より、文法サイズが $\Omega(m\alpha(m))$ であることが示せる！

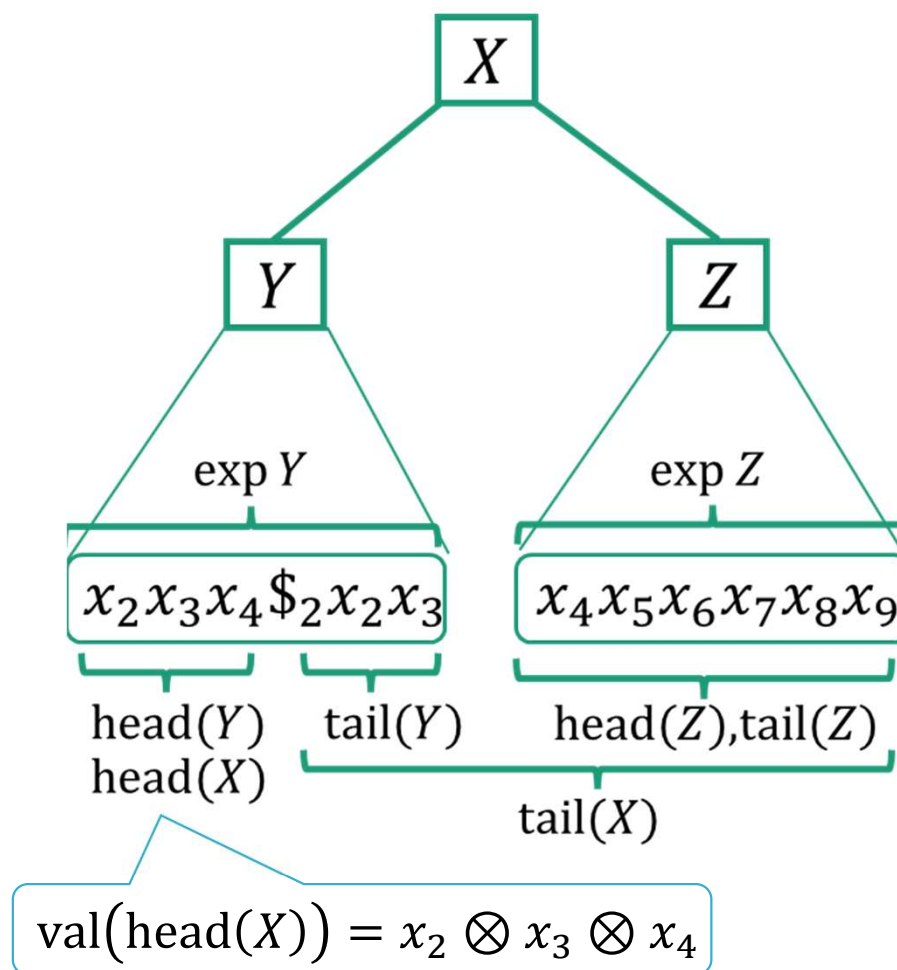
最小文法のサイズ下界 (4)

文法の各非終端記号 X について以下を定義:

- $\text{exp } X$: X が表す文字列
- $\text{head}(X)$: $\text{exp } X$ 中で最初の区切り文字 $\$_i$ の直前までの接頭辞
- $\text{tail}(X)$: $\text{exp } X$ 中で最後の区切り文字 $\$_i$ の直後からの接尾辞

また、 X 上の文字列 $S = s_1 \cdots s_{|S|}$ に対し、

$\text{val}(S) = \bigotimes_{i=1}^{|S|} s_i$ とする



最小文法のサイズ下界 (5)

性質1

変換規則 $X \rightarrow YZ$ について、

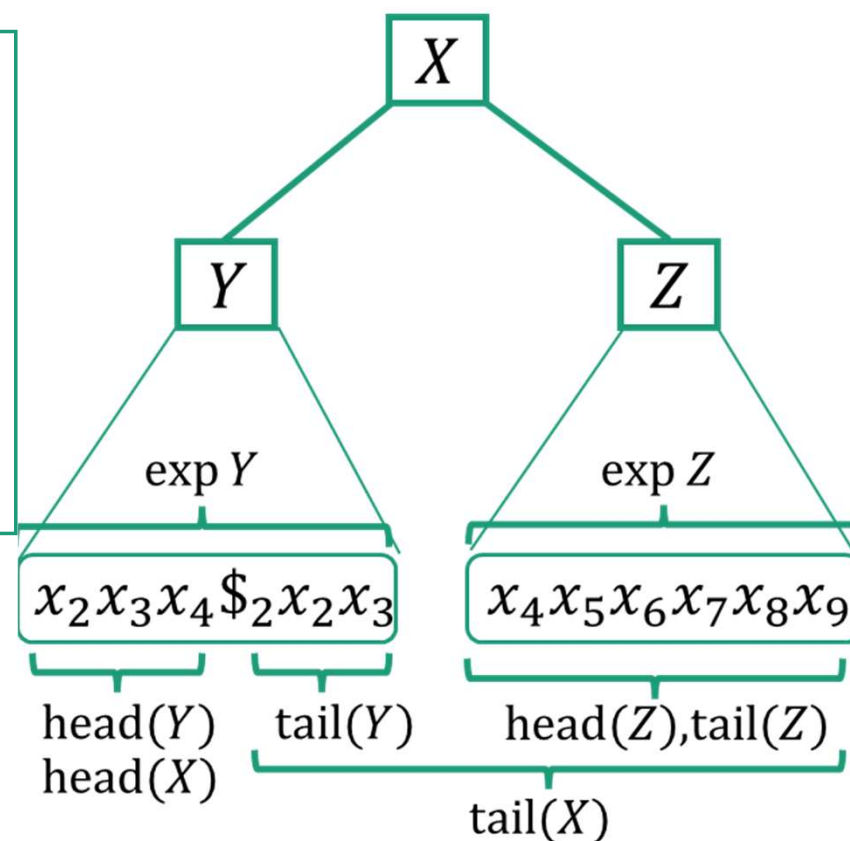
$\text{head}(Y), \text{tail}(Y), \text{head}(Z), \text{tail}(Z)$

の val の値が既知であれば、定数回の半群演算で $\text{val}(\text{head}(X)), \text{val}(\text{tail}(X))$ を計算できる

右図の例:

■ $\text{val}(\text{head}(X)) = \text{val}(\text{head}(Y))$

■ $\text{val}(\text{tail}(X)) = \text{val}(\text{tail}(Y)) \otimes \text{val}(\text{tail}(Z))$



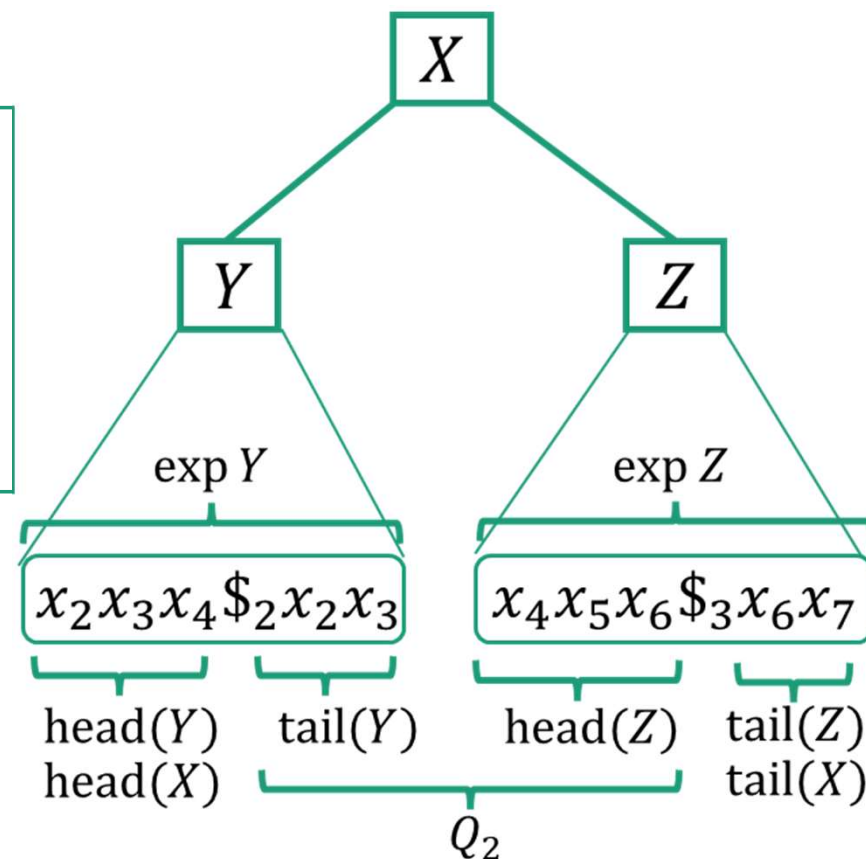
最小文法のサイズ下界 (6)

性質2

変換規則 $X \rightarrow YZ$ について、
 $\text{exp } Y$ に $\$_i$ が、 $\text{exp } Z$ に $\$_{i+1}$ が出現する場合、
 $\text{val}(Q_i) = \text{val}(\text{tail}(Y)) \otimes \text{val}(\text{head}(Z))$ となる

右図の例:

$$\begin{aligned} \text{val}(Q_2) &= \text{val}(x_2x_3x_4x_5x_6) \\ &= \text{val}(x_2x_3) \otimes \text{val}(x_4x_5x_6) \\ &= \text{val}(\text{tail}(Y)) \otimes \text{val}(\text{head}(Z)) \end{aligned}$$



最小文法のサイズ下界 (7)

1. 与えられた文法をチョムスキー標準形に変換する
 - 変換後の文法のサイズは $O(g)$ のまま
 2. 性質1を用いて、各非終端記号に対して $\text{val}(\text{head}(X)), \text{val}(\text{tail}(X))$ を計算
 - $O(g)$ 回の半群演算によって計算可能
 3. 性質2を用いて、 $\text{val}(\text{head}(X)), \text{val}(\text{tail}(X))$ の値から各 $q_i = \text{val}(Q_i)$ を計算
 - $O(m) \subseteq O(g)$ 回の半群演算によって計算可能
- $O(g)$ 回の半群演算で問題の答えを計算できた！

最小文法サイズ vs LZSE分解サイズ

■ 文法のサイズ:

文法サイズ g に対して $O(g)$ 回の半群演算で問題が解ける

一方で、この問題の半群演算回数下界は $\Omega(m\alpha(m))$

→ 最小文法のサイズが $\Omega(m\alpha(m))$ となる入力が存在

■ LZSE分解のサイズ: 下図のような分解で貪欲法でも $O(m)$

$$T = \boxed{x_1} \cdots \boxed{x_m \$1 Q_1 \$2 Q_2 \$3} \cdots \boxed{\$m Q_m \$m+1}$$

→ 最小文法・貪欲LZSE分解のサイズが**オーダーレベルで異なる**！

発表のまとめ

■ 研究の成果: LZ圧縮の制約付き版であるLZSE圧縮を提案

- LZ圧縮～文法圧縮の中間に位置し、文法圧縮より強い圧縮性能を持つ
- 文法圧縮と同じ速度で圧縮領域でのデータアクセスを実現

■ 今後考えたい問題:

- LZSEの圧縮索引を文法圧縮の索引ぐらい高機能にできるか？
 - ▲ 本質的な難しさがあるかもしれないし、実は全部LZSE索引でもできるかも
- 他の圧縮表現との理論的な差は？

