Re-pairアルゴリズムを用いた効率よいVF符号

吉田諭史，喜田拓也*

*北海道大学大学院情報科学研究科コンピュータサイエンス専攻

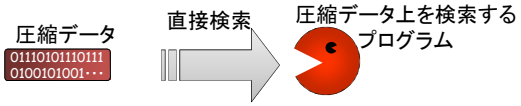
1. 背景

VF符号

入力データを可変長の文字列に分割して，それぞれ固定長の符号語に変換する符号化手法．

特長：圧縮データが取り扱いやすい

⇨情報検索，データマイニングが高速に



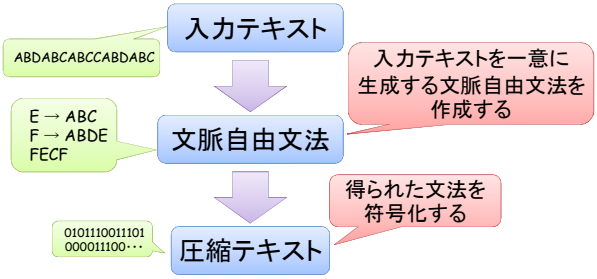
		入力テキスト（情報源記号）	
		固定長 (F)	可変長 (V)
圧縮テキスト（符号語）	固定長 (F)	**FF符号** 等長符号	**VF符号** Tunstall符号
	可変長 (V)	**FV符号** Huffman符号	**VV符号** LZ符号

目的

高い圧縮率を達成するVF符号の開発

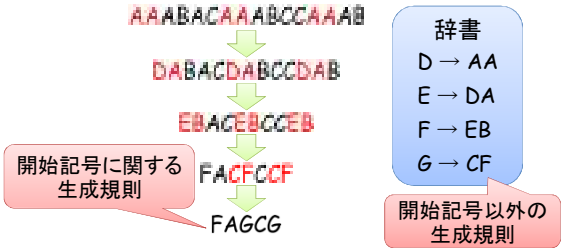
2. 文法変換に基づく圧縮

入力テキストを一意に生成する文脈自由文法を作成し，その文法を符号化する．



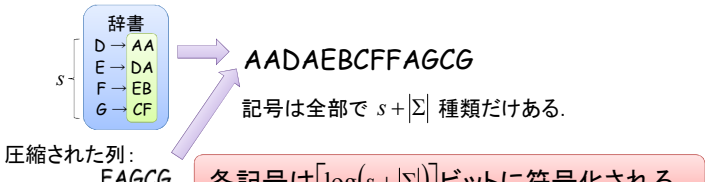
3. Re-pairアルゴリズム [Larsson & Moffat, 1999]

すべてのbigramがuniqueになるまで，最頻出のbigramをひとつの記号に置き換える．



4. Re-pairアルゴリズムを用いたVF符号

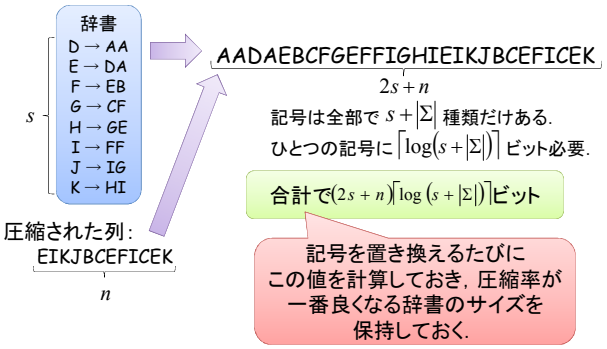
各記号を固定長の符号語で符号化する．



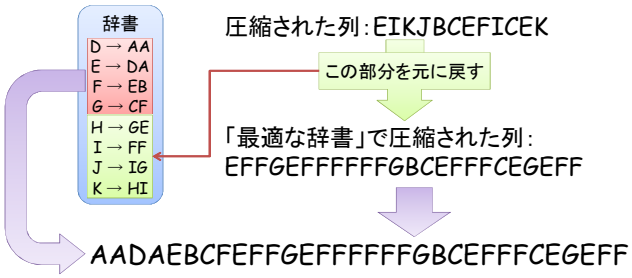
5. 提案手法

記号の置き換えることにより，圧縮率が必ずしも向上するとは限らない．

⇨圧縮率が最良のところで符号化したい．



符号化の際に，「最適な辞書」から外れた部分を元に戻す．



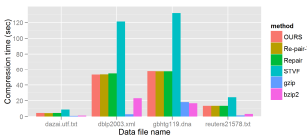
6. 実験

各手法について，圧縮率と，圧縮時間，伸長時間を比較した．

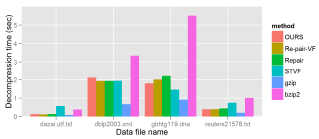
データ

- 日本語テキスト (dazai.utf.txt, 7MB)
- XML文書 (dblp2003.xml, 90MB)
- DNA配列 (gbhtg119.dna, 87MB)
- 英文テキスト (reuters21578.txt, 18MB)

圧縮率の比較．



圧縮時間の比較．



伸長時間の比較．

結果

- 圧縮率は，自然言語について，gzipよりも良い結果が得られた．
- 圧縮速度は，オリジナルのRe-pairとほとんど変わらない結果となった．
- 伸長速度は，bzip2よりも良い結果が得られた．

7. まとめと今後の課題

本ポスターでは，Re-pairアルゴリズムを用いたVF符号と，圧縮率を向上する方法を提案した．また，圧縮率および，圧縮時間，伸長速度を実験的に評価した．他の文法変換に基づく圧縮を取り入れたVF符号を開発することが今後の課題である．