Space-Efficient Re-Pair Compression

Re-Pair [5] is an effective grammar-based compression scheme achieving strong compression rates in practice. Let \( n, \sigma, \) and \( d \) be the text length, alphabet size, and dictionary size of the final grammar, respectively. In their original paper, the authors show how to compute the Re-Pair grammar in expected linear time and \( 5n + 4d^2 + 4d + \sqrt{n} \) words of working space on top of the text. In this work, we propose two algorithms improving on the space of their original solution. Our model assumes a memory word of \( \lfloor \log_2 n \rfloor \) bits and a re-writable input text composed by \( n \) such words. Our first algorithm runs in expected \( O(n/\varepsilon) \) time and uses \((1+\varepsilon)n+\sqrt{n} \) words of space on top of the text for any parameter \( 0 < \varepsilon \leq 1 \) chosen in advance. Our second algorithm runs in expected \( O(n \log n) \) time and improves the space to \( n + \sqrt{n} \) words.