

Byte Pair Encoding (Sennrich et al., 2016)

- Given a dictionary of token types and frequencies.
 - Replace the most frequent pair of characters with a `new unit`. (Record this “merge” operation.)
 - Repeat until the desired number of merge operations is reached.

Current vocabulary	The new merge
low er low e st new er widest	we → <code>we</code>
lo we r lo we st ne we r widest	<code>we</code> r → <code>wer</code>
lo wer lo we st ne wer widest	st → <code>st</code>