

Byte Pair Encoding (Sennrich et al., 2016)

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

Current vocabulary

low**er** low**est** new**er** widest

lo**we****r** lo**we**st ne**we****r** widest

The new merge

we → `we`

`we`r → `wer`