Byte Pair Encoding (Sennrich et al., 2016)

- Given a dictionary of token types and frequences.
  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	The new merge
lower lowest newer widest	