"Neural Machine Translation of Rare Words with Subword Units"

challenge → translating rare & unknown words in NMT systems. Standard NMT models typically operates on a fixed 30K to 50K words, → giving way to open vocabulary problems & facing OOV (out of vocab)

BPE algorithm → to tokenize for better translation
↳ many "words classes" can be translated via smaller units than full words.

BPE. Algorithm

1. Initialization → symbol vocabulary initialized with character vocab. Special end of word symbol "." added to each word.
2. Frequency Calculation → All adjacent symbol pairs in the training data are counted. Weighted according to frequency.
3. Identify most frequent pair of symbols.
4. Merging → each occurence of the most frequent pair is replaced with a new symbol which represents the merged sequence of characters.
5. Each merge operation produces a new symbol → added to symbol vocabulary.
6. Steps 2–5 repeated iteratively.

for better efficiency → pairs crossing word boundaries are not considered. algorithm runs on dictionary with word frequencies.