Neural Machine Translation of rare words with subword units

Neural Networks usually make translation predictions on a fixed output vocabulary of words.

Back-off to an external dictionary look-up might fail - due to - lack of 1 to 1 correspondence between source and target words.
i.e. external dictionary backoff may be bad. Instead a learnt dictionary - from sequence matching might be more helpful to us.

Authors find that subword models preform better for the translation of rarewords than large-vocabulary models and back-off dictionaries, and are able to productively generate new words that were not seen at training time.

Q. How are new words being generated?? If model has not seen a word -> how is predicting possible ? Even when we predict output subwords.

A. Suppose network learnt subwords fly and ing. It can combine them to predict flying which would not have been possible if flying was not explicitly present in the dictionary

Append each word with a end_of_word symbol and separate the words into characters.
Now successively merge 'subwords' frequently occurring together.

Initially merge characters occurring together. Then pair of words occurring together and so on.
How long we continue to do this is a hyperparameter.

Two methods to apply BPE are proposed:

- Apply BPE separately on the source and target sentences.
    - Named entities may be segmented differently in both languages

and make it diff

for Network to learn a mapping

- Combine the source and target sentence and learn and joint BPE subword model
    - Transliteration required to bring both the corpora in the same script
    - Translated sentences have to be then retranslated in the original script