Machine Translation

Exercise 05

Wu Mengjie, matriculation number: 22-747-869
Xu Ruomeng, matriculation number: 22-748-339

1 Experiments with Byte Pair Encoding

The BLEU scores of the three experiments are as follows :

	use BPE	vocabulary size	BLEU
(a)	no	2000	14.7
(b)	yes	2000	22.7
(c)	yes	5000	23.9

Investigate the difference in BLEU on the test set:

In experiment (a), the model is on the word-level and does not employ Byte Pair Encoding (BPE). This approach is restricted in handling unknown or rare words, typically resulting in a lower BLEU score. The model struggles to effectively manage words not contained within the fixed vocabulary.

The experiment (b) uses the same vocabulary size as that of the experiment (a). But incorporating BPE allows the model greater flexibility in managing unseen words and vocabulary variations. BPE enables the breakdown of words into smaller, reusable segments, i.e. subwords, significantly improving model translation quality by better adapting to variations in input data, thus achieving a higher BLEU score than that of the experiment (a).

In experiment (c), by further increasing the BPE vocabulary size, the model can learn and utilize the nuances of language more detailedly. A larger vocabulary size allows for a more precise representation and translation of texts, resulting in an improvement in the BLEU score, indicating that increasing vocabulary size can enhance translation accuracy under certain conditions.

The comparison of the differences in BLEU on the test set demonstrates that the use of BPE significantly enhances the performance of translation models. Particularly under vocabulary constraints, the BPE approach effectively improves the model's capability to handle complex and variable language inputs. Moreover, appropriately increasing the BPE vocabulary size can further enhance translation quality, although the magnitude of this improvement might not be as substantial as the initial introduction of BPE itself. These findings emphasize the importance of choosing the right vocabulary handling strategy when designing machine translation models.

To look at the translations of three experiments manually and describe how translations differ:

Analyzing the translation results from the three experiments, we see distinct variations in the quality and coherence of the translations:

For experiment (a), extensive use of the placeholder ‘<unk>’ (unknown token) indicates the model's frequent encounters with words not in its vocabulary. The overall coherence is severely compromised, with fragmented sentences and repeated use of ‘<unk>’, making the translation difficult to understand and follow. This translation lacks clarity and fidelity, reflecting the model's limitations in handling a wider range of vocabulary with only full words.

For experiment (b), there is no use of ‘<unk>’ in the translation, suggesting better handling of unknown or rare words due to the BPE technique. The translation is more coherent and fluid compared to experiment (a), although there are still some nonsensical phrases and minor errors. The use of BPE allows the model to piece together more accurate phrases and sentences, improving readability and context understanding.

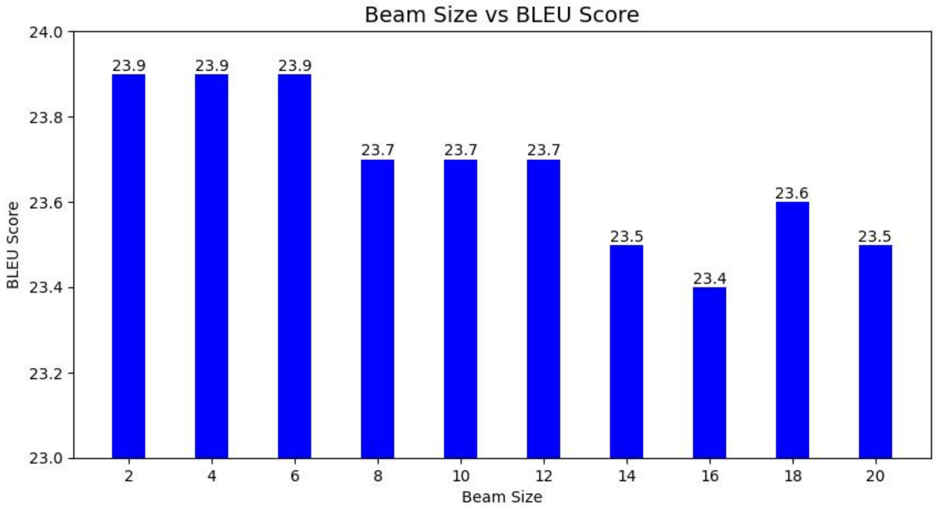
The translation from experiment (c) has the best coherence and readability among the three, with a more extensive vocabulary allowing for finer nuances. Improved translation accuracy and completeness, with a clearer conveyance of ideas and more structured sentences. Despite occasional odd word choices and phrasings, this translation demonstrates a significant enhancement in handling complex language constructs.

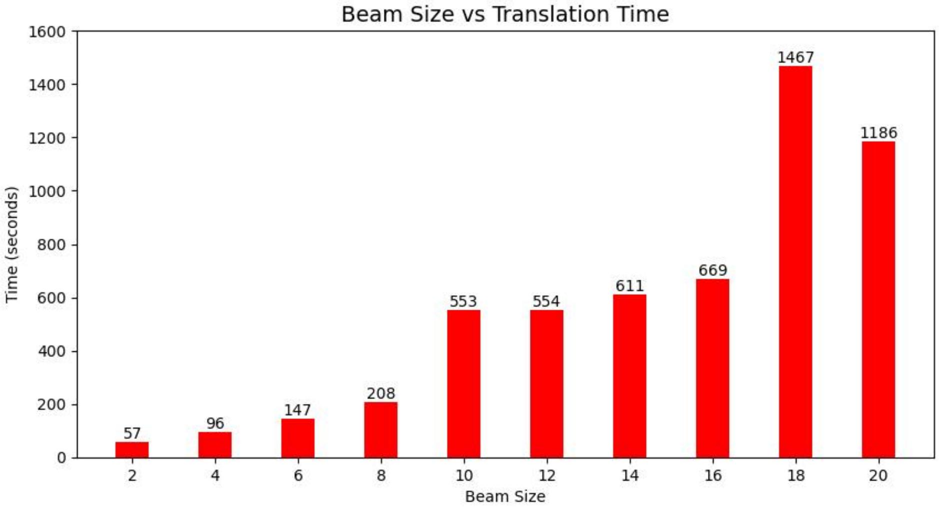
The transition from Experiment (a) to (c) shows a clear upward trajectory in translation quality. The introduction of BPE greatly reduces vocabulary limitations by breaking down words into subword units, enabling the model to manage unknown words more effectively. Increasing the vocabulary size in experiment (c) further refines this advantage, allowing the model to capture and translate more complex ideas and technical terms with higher accuracy. The results underscore the importance of subword segmentation such as BPE in enhancing machine translation output, particularly when dealing with languages that have rich morphology or where the direct translation of whole words may be insufficient due to a small fixed vocabulary.

Overall, these experiments illustrate how the strategic use of BPE and the adjustment of vocabulary size can significantly impact the quality of machine translation, particularly in terms of handling linguistic diversity and improving the model's ability to deliver coherent and contextually accurate translations.

2 Impact of beam size on translation quality

For Task 2, we use transformer_c as the model because it has the highest BLEU score in Task 1. The following two graphs are plotted to visualize the impact of beam size on the BLEU score and the impact on time taken to generate the translations.





The graph 'Beam Size vs BLEU Score' depicts a more subtle impact of beam size on BLEU scores, which are used to measure the accuracy and quality of translations. Initially, the BLEU scores remain quite stable around 23.9 for beam sizes 2 to 6. A slight decline begins from beam size 8 onwards, reaching the lowest points at beam size 16. The decrease in BLEU scores is modest but noticeable, indicating that increasing the beam size does not necessarily improve translation quality.

The graph ‘Beam Size vs Translation Time’ shows a sharp increase in translation time as beam size increases. Starting from a relatively modest 57 seconds for a beam size of 2, the time required escalates dramatically at the beam size of 10 and 18, even reaching 1,467 seconds for a beam size of 18. Although there is a slight drop to 1,186 seconds at a beam size of 20, the overall trend suggests that larger beam sizes require significantly more computational time.

Given the data from the graphs, our choice of beam size would be driven by considering a balance between computational efficiency and translation quality. We would choose the beam size of 6 in the future.

According to the graphs, the BLEU scores for beam sizes 2, 4, and 6 are all maintained at 23.9. Although a beam size of 2 is more time-efficient, choosing a beam size of 6 allows for maintaining high BLEU scores while increasing the model's adaptability and robustness across different texts and complex scenarios. This balance is based on sustaining translation quality while accepting a moderate increase in time. Thus, a beam size of 6 represents a practical compromise between high-quality translation and computational efficiency.