**Experimental Settings**

For this experiment, because of time and memory limitation, I had to use **English-German pair from** **Multi30k with 30,000 sentences** (Torchtext native library) for default experiment dataset. The paper utilized WMT14 dataset in English-French pair with 12,000,000 sentences (160,000 source words / 80,000 target words). To compare the performance in the same language, I also used **ANKI English-French pair with 120,000 sentences** (13,000 source words / 9,000 target words), smaller dataset, yet large enough to provide reasonable performance.

Regarding the architecture of the model, the paper used 4 layers for LSTM, 1000 hidden states, 1000 embedding dimensions, uniformly distributed initialization (between -0.08 and 0.08) for all the parameters, adaptive learning rate, and 128 batch size. I did keep the initialization. For other configurations, I changed in order to conduct efficient experiment under limited resources. My configurations are in the result table below. For the performance metrics, I used BLEU score and output sentence comparisons. Further, to have more semantic understanding without using additional translation, I set target language as English. By tweaking **different layer sizes, reversing the input texts, and adding dropout layer**, I experimented to see various performance comparisons.

**Results**

As you may see from the result tables, BLEU scores for Multi30k-EN-DE were very high, over 30 points score except for when the LSTM has shallow architecture. Simply, our models showed better results. However, I found out that English-German pairs (for Multi30k) tend to perform better on BLEU scores; thus, current performance should be discounted. Then, for the second dataset, where I used ANKI EN-FR, the result was very similar to the paper, even though total size of the dataset was much smaller. Therefore, I can say that my BLEU scores are on-par with the paper.

In fact, actual output sentences are worthy to be discussed. As you can see in the table below, I have picked three random output sentences from two of my best performing models, for each EN-DE and EN-FR cases. The results demonstrate that overall translation is fairly well. However, typically, **subjects of the sentences are well translated** while either **object or actions (in verbs, adverbs, or adjectives) mis-translated.** Yet, when it is incorrect, the meaning is somewhat preserved. For instance, **when target was “green vegetable”, the model produced “tree”,** which is not exact in the meaning, but it does convey some meaning (as vegetable can be subset of bigger tree group). Moreover, **certain numerics or quantities, such as “two”, “a man”, “men” or “three” are also well preserved**. This is probably because of the order and simplicity in meanings of these words.

**Discussion**

Because dataset was different, the results cannot be generalized right away. Nonetheless, I believe that overall performance in BLEU score and output sentences both demonstrate that the experiment was on-par with the results described in paper. We have learned that Dropout enables efficiency in training and improvements in performance. Thus, I tried to incorporate Dropout layer in LSTM of my seq2seq model. However, for some reason, both training efficiency and performance were benefited from it. Although I did not include the performance here, this is why my code contains both normal and with dropout models.