Task 1:

Source language: Italian
Target language: German

The task here is to create a model that can reliably translate documents from Italian to German. Three different models have been created, listed below:

| | use BPE | vocabulary size | BLEU |
|---|---|---|---|
| (a) | no | 2000 | 6.5 |
| (b) | yes | 2000 | 10.7 |
| (c) | yes | 1000 | 4.3 |

There is one additional script which I have called transcription.py that sub-samples the data (if necessary) and tokenizes it.

In general, one could argue that the BPE model trained on 2000 “words” is superior to the word-level model as the BLEU scores here clearly indicate. The BPE model trained on 2000 words is better = gets a better BLEU score than the one trained on 1000. My impression here is that the word-level model has the advantage of not producing any new words. However, it also requires a lot more vocabulary to make up for the losses in flexibility.

When looking at the translations made by a BPE model, I could see certain anomalies that are not present in the translations made by a word-level model. These include disfigured words that are composed by parts that do not belong together. These include such words as “Peträumt” and “Spitzenium”. Comparing the BPE model trained on a vocabulary size of 2000 compared to one trained on a vocabulary size of 1000 quickly reveals that the former translation approximates some of the original testing sentences, whereas the latter mostly produces non-sense.

The word-level model does have the advantage of not producing any new words. But many of the words, which happen to be the most important ones in terms of information content, are completely ignored and simply represented by <unk>. The word-level model needs a big vocabulary file to make up for the losses that BPE models can more easily adapt to.

Task 2:

I have used translations made by the BPE model trained on a vocabulary size of 2000 for the second task.

We can observe that there is a small but significant change in the BLEU score when beam size is changed. We get the best BLEU scores for beam size = 2. The lowest ones are found for beam size = 0 and 10. There is a gradual linear decline from beam size = 2 to = 10, whereas a change from beam size = 1 to = 2 significantly improves BLEU scores. Considering this example, I would argue that a medium-low BLEU score (2 or 3) would be the ideal setting for BPE models.

| Beam size | BLEU |
|---|---|
| 0 | 10.7 |
| 1 | 10.7 |
| 2 | 11.4 |

| | |
|---|---|
| 3 | 11.3 |
| 4 | 11.2 |
| 5 | 11.2 |
| 6 | 11 |
| 7 | 11 |
| 8 | 11 |
| 10 | 10.7 |