AI539 Spring 2024 Homework IV

Rigved Naukarkar – naukarkr@oregonstate.edu

June 1, 2024

Task 1.1

For output a to equal the v_j, the q should be large such that its dot product with k_j is large. The dot product of q with other k's should be zero or at least very small compared to k_j.

Task 1.2

q vector will be like: $q = \beta(k_a + k_b)$. Here β will be a large constant. So both α_a and α_b will be large, and equal, and for all other i's α will be small.

Task 1.3

Now, q will be set like

$$\begin{aligned} q &= c(\mu_a\lambda_a + \mu_b\lambda_b) \\ qk_a^T &= c(\mu_a\lambda_a + \mu_b\lambda_b)k_a^T \\ &= c(\mu_a\lambda_a + \mu_b\lambda_b)\mu_a^T\lambda_a \\ &= c((1)\lambda_a\lambda_a + 0) \quad \text{as } \mu_a^T \cdot \mu_a = 1 \text{ and } \mu_b \cdot \mu_a = 0 \end{aligned}$$

Similarly, we can get $qk_b^T = c\lambda_b^2$. All the other qk_i^T's will be equal to 0 as $\mu_a \cdot \mu_i = 0$ and $\mu_b \cdot \mu_i = 0$

If λ_a and λ_b are the same, then the α values will be the same as in Task 1.2. λ_i's are sampled randomly, so the values will probably not be the same, although q will be dominated by α_a and α_b values as we have a large c.

Task 1.4

It is given that $a = \frac{1}{2}(a_1 + a_2)$. So for, $a \approx \frac{1}{2}(v_a + v_b)$, we need to make $\alpha_a \approx 1$ and $\alpha_b \approx 1$.

If we have keys like in Task 1.3, then we will have

$$\begin{aligned} q_1 &= c(\mu_a \lambda_a) \\ q_1 k_a^T &= c(\mu_a \lambda_a) k_a^T \\ &= c(\mu_a \lambda_a) \mu_a^T \lambda_a \\ &= c((1)\lambda_a \lambda_a) \quad \text{as } \mu_a^T \cdot \mu_a = 1 \end{aligned}$$

The q_2 will be $q_1 = c(\mu_a \lambda_a)$ and $q_1 k_b^T$ will be calculated like above.

The a_1 is only dependent on the α_a, which depends on $q_1 k_a^T$. So to maximize the α_a we can adjust the value of c to be large. α_b will be calculated the same way. Then we will get the desired result.

Task 2.1

```
2024-06-01 02:44:44 INFO     | Test Loss: 2.349 | Test PPL:  10.476 | Test BLEU 33.54
```

Figure 1: BLEU score and Perplexity scores

Task 2.2

Please find the examples of the generated attention graphs. Also the patterns observed in the captions.

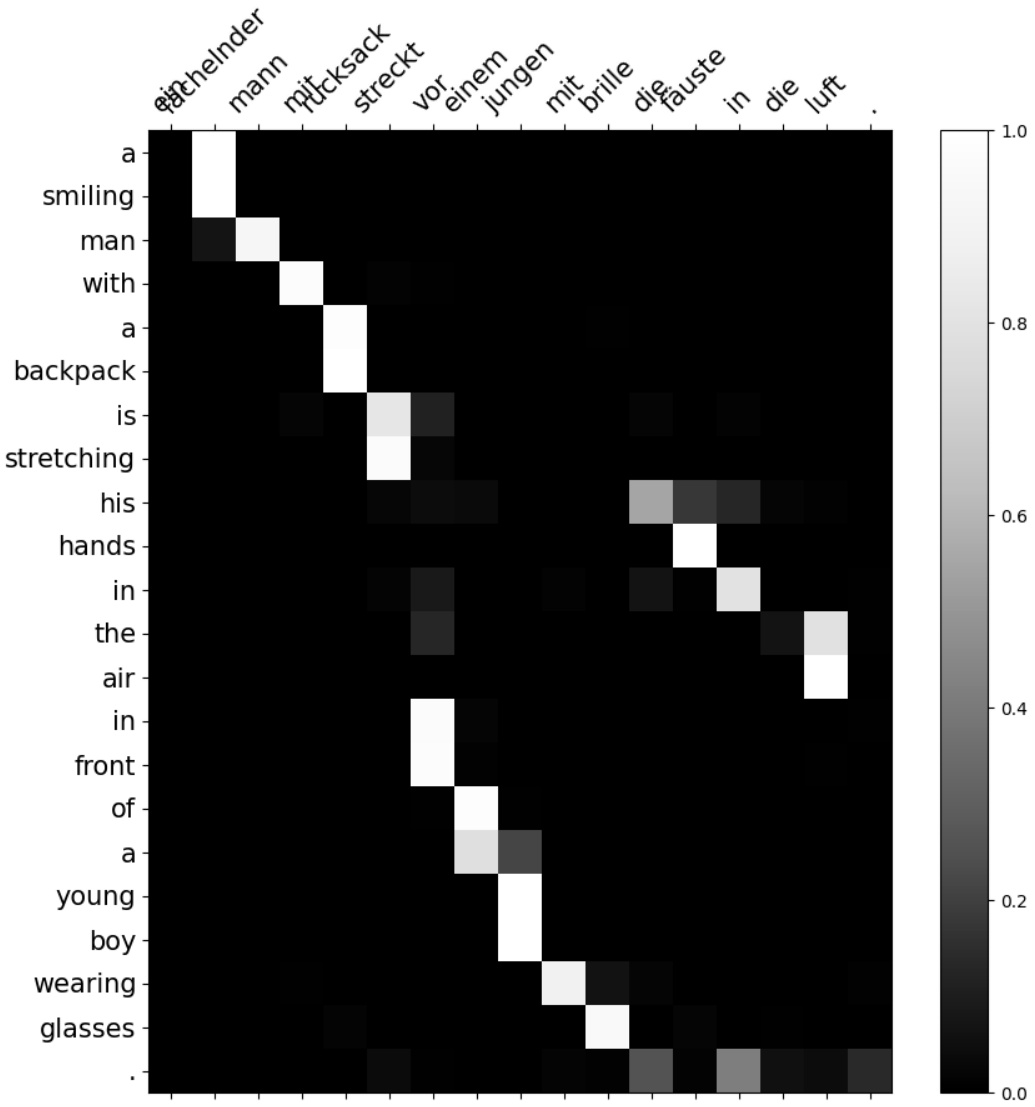

Figure 2: Correct translation. The subject-object-verb pattern in German is being translated into subject-verb-object in English.

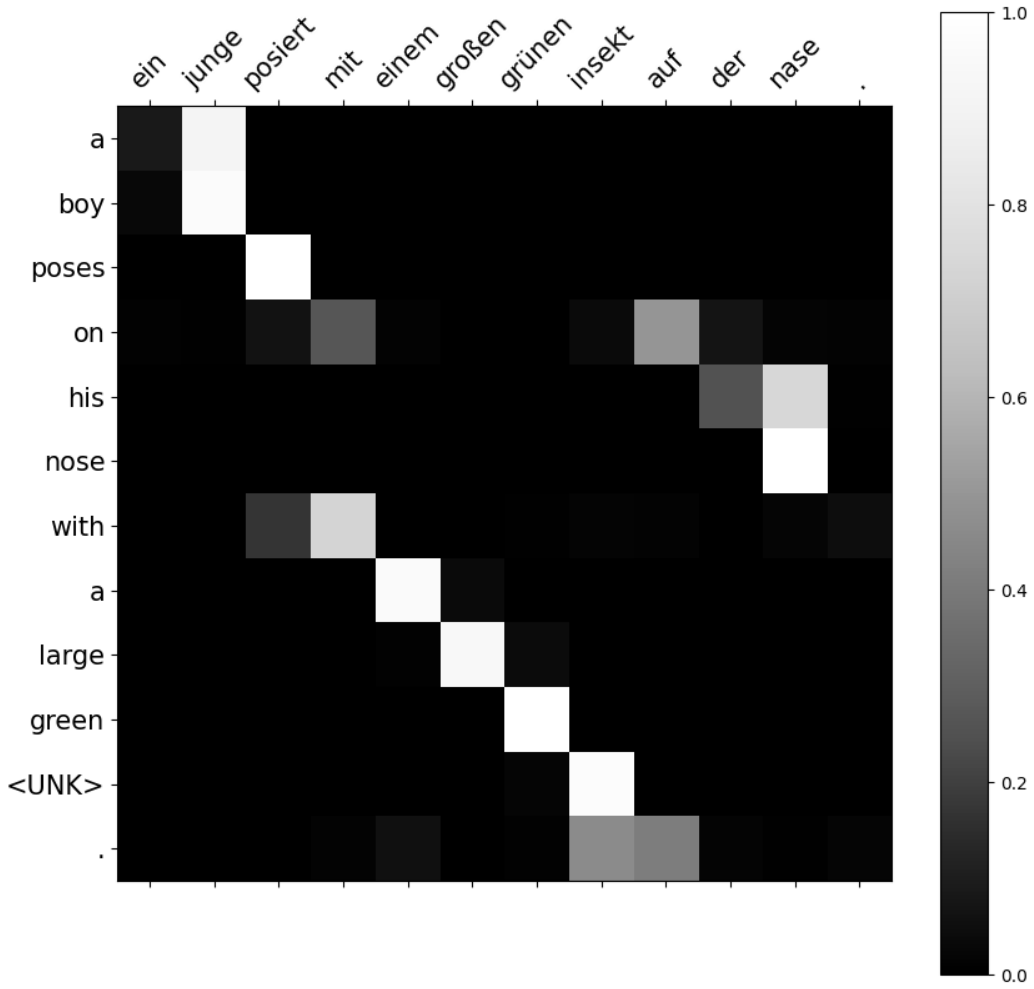

Figure 3: The word insect is not in the vocabulary. The subject-object-verb pattern in German is being translated into subject-verb-object in English.

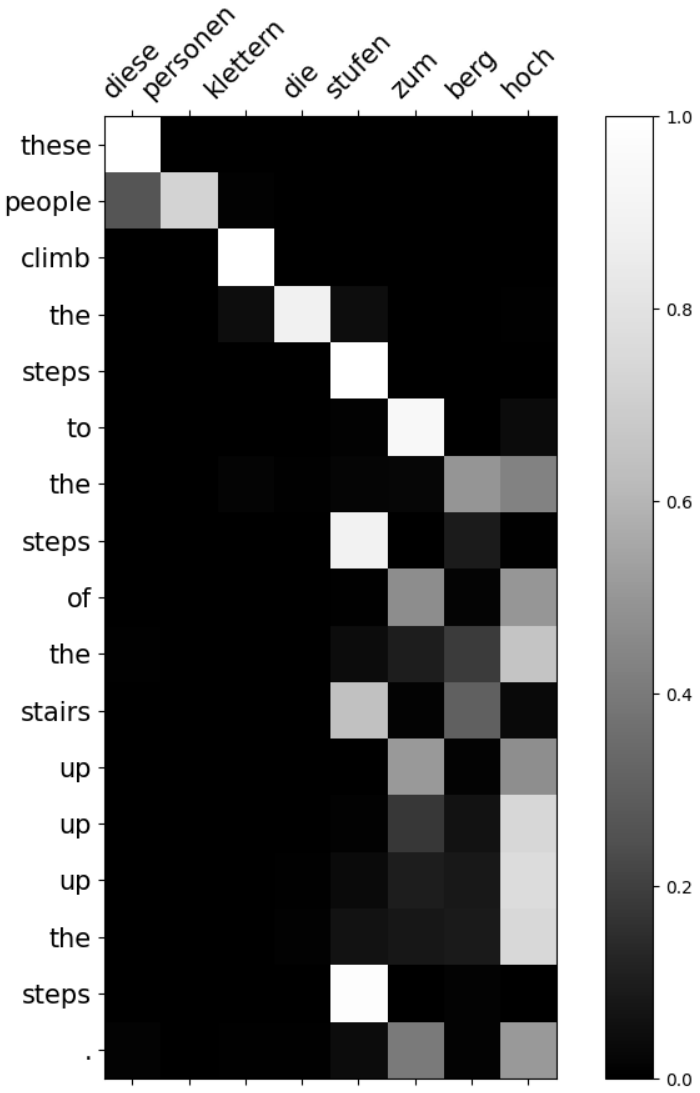

Figure 4: This translation seems wrong. The words 'zum', 'hoch' seem to come from many different words. One German word can carry the meaning of multiple words in English.

Task 2.3

1. Dummy
   (a) PPL Mean: 18.18
   (b) PPL Variance: 0.05

(c) BLEU Mean: 16.21

(d) BLEU Variance: 0.28

2. Mean Pooling

(a) PPL Mean: 15.79

(b) PPL Variance: 0.014

(c) BLEU Mean: 18.15

(d) BLEU Variance: 0.033

3. SDP

(a) PPL Mean: 10.57

(b) PPL Variance: 0.023

(c) BLEU Mean: 33.36

(d) BLEU Variance: 1.18

The BLEU scores seem to increase in this order: Dummy, Mean Pooling, and SDP, ranging between 15.71 and 34.06. The PPL values on the other hand are decreasing as we go through Dummy, Mean poling, and SDP. Observed a negligible variance in the values of BLEU and Perplexity. Overall, the SDP mechanism seems the best for these metrics. Please find the screenshots below.

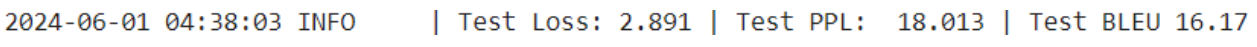

Figure 5: Dummy() Run 1 Scores

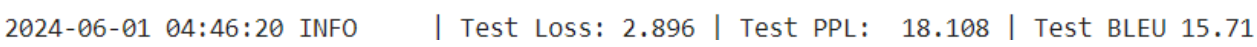

Figure 6: Dummy() Run 2 Scores

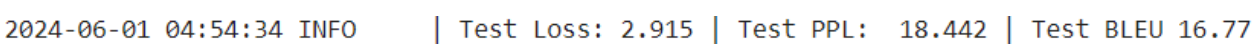

Figure 7: Dummy() Run 3 Scores

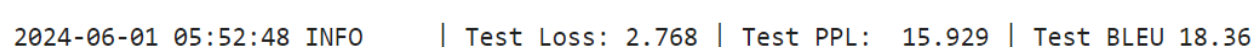

Figure 8: Mean pool Run 1 Score

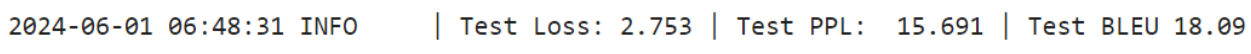

Figure 10: Mean pool Run 3 Score

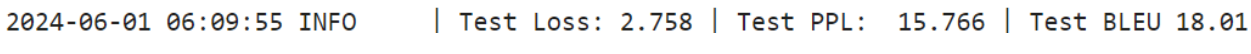

Figure 9: Mean pool Run 2 Score

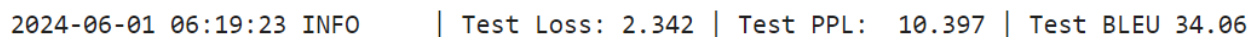

Figure 11: SDP Run 1 scores

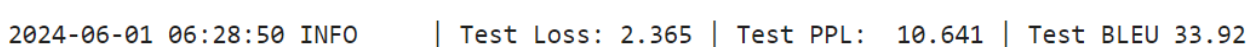

Figure 12: SDP Run 2 scores

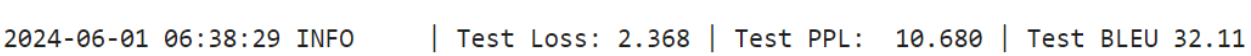

Figure 13: SDP Run 3 scores