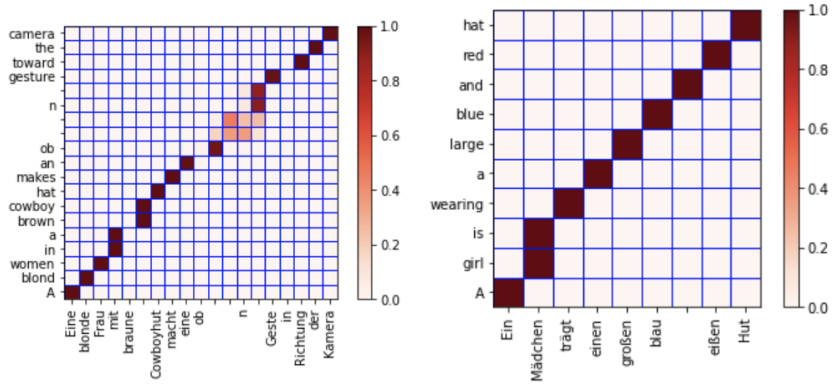
Machine Translation: Attention Visualization

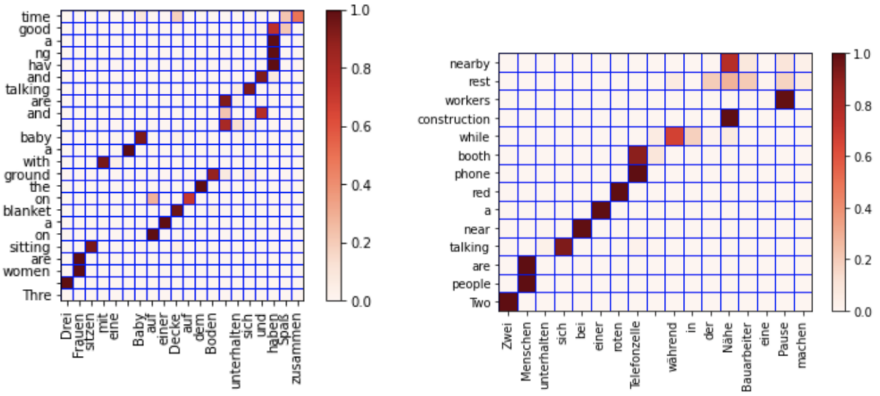
NOPPAPON CHALERMCHOCKCHAROENKIT

In the context of our Neural Machine Translation project, a crucial undertaking involves the integration of an attention mechanism within the decoder of our seq2seq model. Below, we delve into a deeper understanding of the learned attention by examining attention map plots for four English-German sentence pairs drawn from the validation dataset. Additionally, the concluding section of this report will present the outcomes of our IBM1 alignment model, trained on a set of 1000 English-German sentence pairs sourced from the Multi30k dataset, and provide a comparative analysis with our attention-based approach..

1. ATTENTION



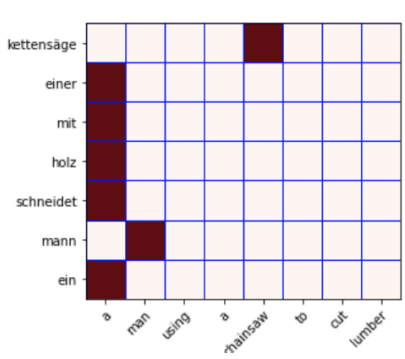
In the two plots displayed above, it is evident that, in most instances, the attention mechanism consistently associates each source input word with a word in the target sequence with a high probability (one-to-one), as indicated by the dark red tiles. Furthermore, both plots exhibit a clear, straight diagonal line, suggesting that in these sentences, the word order between their English and German versions closely aligns. This appears to be a common trend for the English-German sentence pairs in the validation set.



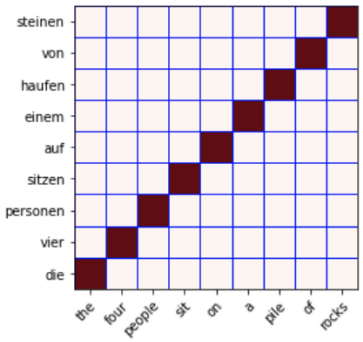
In the left plot, although there are segments with distinct diagonal lines, they appear somewhat fragmented, indicating variations in sentence structure between the English and German versions. Furthermore, several 'many-to-one' mappings are observed, such as 'hav', 'ng', and 'a', all mapping to 'haben', which translates to 'have' in German. This is unsurprising, as 'having a' and 'have' essentially convey the same meaning but are used in different tenses and contexts, and the attention mechanism did a great job identifying that.

In the right plot, particularly towards the end, we observe instances of 'many-to-many' relationships. Specifically, we can see how the phrase 'rest nearby' derives its meaning from 'in der Nähe Bauarbeiter ein pause machen,' which roughly translates to 'Nearby construction workers take a break.' This occurrence demonstrates the capability of the attention mechanism to infer context, with 'rest' capturing the reason for construction workers taking a break."

2. IBM1



(a) IBM1: Poor Result



(b) IBM1: Good Result

IBM1 essentially establishes a mapping from each word in the source sentence to a word in the target sentence that has the highest probability. These probabilities for each English-German word pair are generated during the training process using the training dataset. Consequently, it's possible that certain word pairs do not exist in our probability table. This is exemplified in the left figure, where most word mappings have a probability of 0 and default to 'a'. In contrast, in the right figure, the correct word pair has been encountered during the training process, resulting in a highly accurate mapping.

While not depicted in the figures above, it's worth noting that the IBM1 model can face challenges when dealing with words that have multiple meanings or can be used in various contexts in one language but have specific translations in the other language. Additionally, frequent co-occurrence of two words in the same sentence can introduce confusion into the mapping process.

3. ATTENTION & IBM1 COMPARISON

In the attention model, when it effectively maps words from the source sentence to corresponding words in the target sentence with a high probability, its word alignment visualization closely mirrors that of the IBM1 model. Conversely, as highlighted in the attention section, when words in the source sentence exhibit complex many-to-many relationships with words in the target sentence in terms of context sharing, this intricacy is manifested through the attention probability distribution, which the IBM model cannot depict.