1 Question 1

The sampling techniques outlined in the presentation are :

1. Ancestral sampling.
2. Greedy search.
3. Beam search.

We opted for the greedy decoding strategy which consists of sampling the word that maximizes the log-likelihood conditioned on the previous selected words and the source sentence context. The method is far more efficient computationally as opposed to more exhaustive techniques, which maintain all possible translations and select the one with the best score, or beam search which maintains K hypotheses at a time. The main drawback is that it is heavily sub optimal meaning it generates samples of poor quality comparatively speaking. So a compromise is to be made between quality and complexity.

2 Question 2

One of the major problems in the generated translations is the issue of repetition. we can see that the sentence **'The kids were playing hide and seek'** is translated as '**les enfants jouent cache cache cache cache caché caché caché caché caché caché caché caché caché caché caché caché caché caché caché dentifrice perdre caché risques rapide caché risques éveillés**', and the phrase **'I am a student.'** translates to **'je suis étudiant . . . . . .'**
In the second case it is clear that no word is repeated, the dot marking the end of the sentence is. This problem could be mitigated by tokenizing punctuation such as '.', '?' or '!' which occur at the end of the sentence as ¡EOS¿. Although this does not guarantee to give the expected results.
The first error is that of the repeating words, which can be exactly the same or very similar. The issue here is that the word that is generated next takes into account all the hidden states of the input sequence as well as the last hidden state of the last generated word, hence global attention. An alternative strategy is to infer the weights $\alpha_{t,i}$ only based on the hidden states present within a window, referred to as local attention. This approach is much more practical when translating long sequences, i.e. paragraphs or whole documents, it might also remedy the problem of repetition we have witnessed.

3 Question 3

We can see 4 types of patterns emerge when aligning the the source and target words based on their respective weights.
Firstly we have adjective-noun inversion which is present when translating english to french :

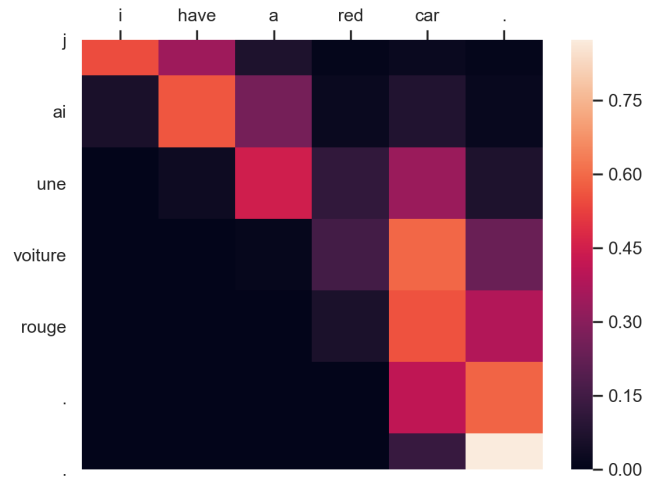

Figure 1: Alignment of the sentence ”I have a red car” with its translation.

We can clearly see that when red is referenced in the source sentence, it barely stimulated the generation of **rouge**, in fact, **voiture** had more weight than **rouge** when the model generated **red**. But when **car** occurs in the original sequence both **voiture** and **rouge** become much more weighty. We can also see that the model chose the article ”une” instead of ”un” thanks to the attention model.

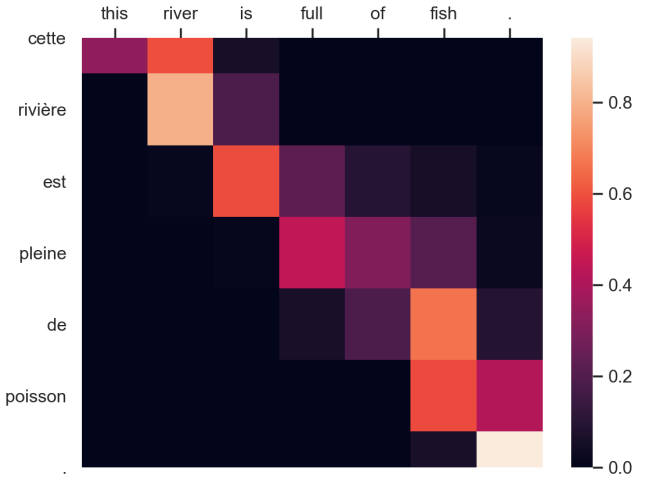

Figure 2: Alignment of the sentence ”This river is full of fish.” with its translation.

In the same spirit, in figure 2, we see that the adjective ’**pleine**’ is feminine as is the noun ’**rivière**’. This property of language is not existent in english (No masculine vs. feminine words).
So the model is able to deduce these patterns of syntax by itself and use them appropriately (In most cases).

Figure 3: Alignment of the sentence ”I did not mean to hurt you” with its translation.

Lastly in 3, we can see that the word **”not”** is solely responsible for generating 3 tokens : **’n’**, **’ai’** and **’pas’**, it is aligned with **’n’** with a score close to 1. The word **’did’** was helpful in predicting **’ai’** since the meaning would have changed if **’did’** was to be replaced with **’should’**, for example. After that the negation word **’pas’** is again almost perfectly aligned with not. Hence the importance of recurrent attention models.

4 Question 4

The translations are as follows:
”She is so mean” = ”elle est tellement méchant méchant .” ”I did not mean to hurt you” = ”je n ai pas voulu intention de blesser blesser blesser blesser blesser blesser .”
Drawbacks outlined in the previous questions aside, we see that the word **”mean”** has two different meanings, the first is a verb meaning to intend, to convey, and the second is an adjective meaning unkind or spiteful. These are homonyms.
The property of language models illustrated is that of contextual representations of words. So here the embedding of the word **”mean”** differs between the two instances based on the neighboring words making the model able to distinguish and correctly predict the right word given its meaning in the relevant context.

References