1 Question 1

Compared with other decoding strategies in NMT systems, the greedy strategy has significant advantages and disadvantages.

1.1 Advantages of greedy decoding

- **Simplicity**: Greedy decoding is simple and requires no complex parameters, making it easy to implement.
- **Computational efficiency**: By choosing the most probable token at each stage, this strategy is computationally efficient. It does not require intensive exploration of the search space, which makes it fast.
- **Performance on certain metrics**: On metrics such as BLEU (Bilingual Evaluation Understudy) and NLL (Negative Log-Likelihood), greedy decoding can sometimes give satisfactory results.

1.2 Disadvantages of greedy decoding

- **Limited quality**: Because of its simplicity, greedy decoding can produce translations of limited quality. It only takes into account the local information at each step, without considering the overall context of the output sequence. This can lead to inconsistencies and errors in translation.
- **Lack of diversity**: Gluttonously generated translations tend to lack diversity. They always follow the most likely path, which means the results can be redundant and repetitive.

2 Question 2

We have identified two major problems in the translations generated by our NMT system:

- **Over-translation**: Some words are repeated and translated several times in the output. This can lead to redundant and unnatural translations. This problem arises because attention, as calculated at each decision stage, does not take into account past translation decisions. As a result, the model may repeatedly choose the same source word for translation, leading to unnecessary repetition.
- **Under-translation**: On the other hand, some source words are not translated at all, or are insufficiently translated. This is also the result of limited attention, as the model may not attach sufficient importance to certain source words to include them in the translation.

To solve these problems, we are considering two solutions:

- **Integrating the notion of coverage**: One possible approach would be to add the notion of coverage to our NMT system. This means that source words that have already contributed significantly to the translation of a target word in the past are assigned a lower weight for subsequent translations. To implement this idea, we could use coverage structures that indicate whether or not each source word has already been translated. This information would be taken into account when calculating alignment probabilities. In this way, we can mitigate over-translation and ensure greater variety in translations.
- **Using local attention instead of global attention**: Currently, our translation system struggles to process long sentences efficiently. By using local attention instead of global attention, we can improve the translation of these sentences. Local attention allows us to focus on smaller parts of the source sentence at each stage, which can be particularly useful for long sentences.

3 Question 3

In the case of literal (word-for-word) translations, the few examples we've observed show that for each word in the target, the model's attention is most focused on the associated word in the source, as can be seen in Figure 1.

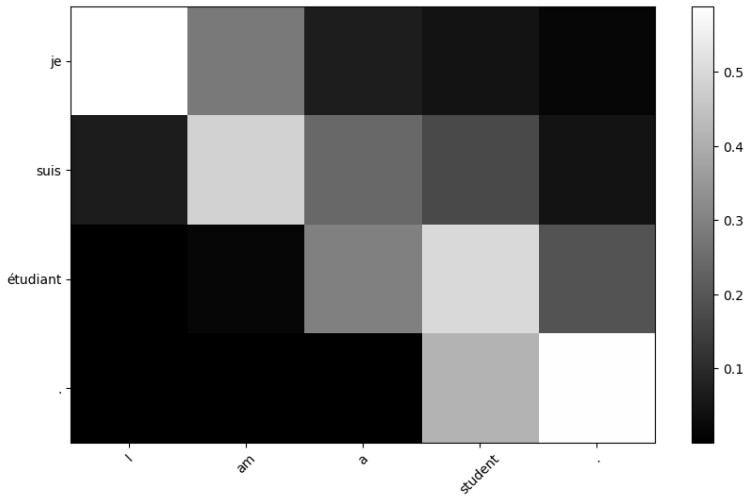

Figure 1: Source/target alignments. Source : *I am a student.* - Target : *Je suis étudiant.*

A first particular case to observe is the inversion of an adjective and a noun during the translation process (for example: *red car - voiture rouge*, illustrated in figure 2). In the example shown here, the model's attention when predicting the noun (*voiture*) is focused more on the corresponding noun in the source (*car*); the same constant is observed for the prediction of the adjective (*rouge*), where attention is also focused more on the same noun and less on the corresponding adjective in the source (*red*), which comes second.

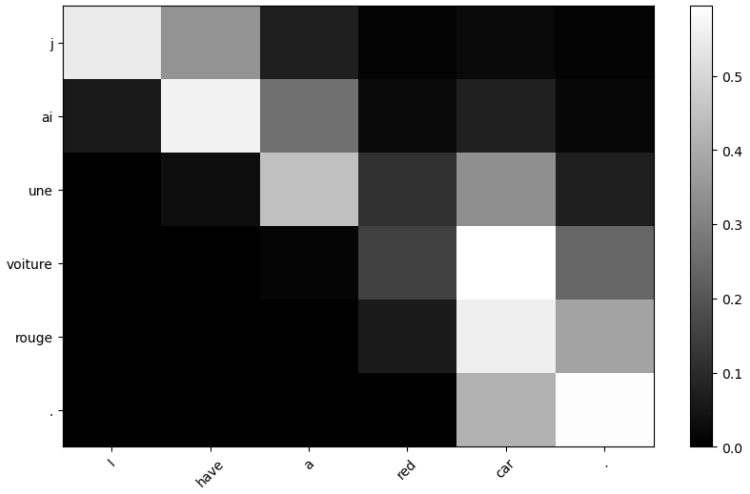

Figure 2: Source/target alignments. Source : *I have a red car.* - Target : *J ai une voiture rouge.*

Another special case is when the translation leaves an article in front of a noun, knowing that the corresponding noun in the source has no determiner. For example: *My brother likes pizza. - Mon frère aime la pizza*, as shown in Figure 3. We notice that for the prediction of the article and the noun (*la pizza)*, the model's attention is most focused on the corresponding noun in the source (*pizza*).

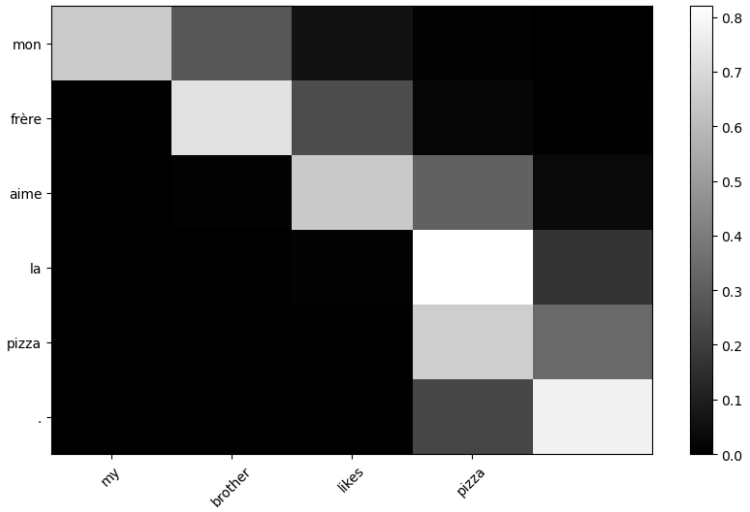

Figure 3: Source/target alignments. Source : *My brother likes pizza.* - Target : *Mon frère aime la pizza.*

4 Question 4

We note a notable difference in the translation of the word *mean* in the two sentences given. In the sentence *I did not mean to hurt you* it is translated as *avoir l'intention de* whereas in the sentence *She is so mean* it is translated as *méchante*.
This translation variation illustrates a fundamental property of language models, namely the notion of context. Language models, especially forward ones, condition the probability of a word (t_k) on previous words from t_1 to t_{k-1}. In this way, the semantics of a word depend closely on its context.

$$P(t_1, t_2, ..., t_N) = P(t_k | t_1, ..., t_{k-1}) \tag{1}$$

In the context of machine translation, this consideration of context is essential for selecting the appropriate meaning of a polysemous word, such as ”mean” in our example.