AMÉDÉE Romain
romain.amedee@telecom-sudparis.eu

Lab session # 1
ALTEGRAD 2023

10/17/23

1 Question 1

By picking the most likely symbol at each step, our greedy search is very computationaly efficient. However, it is also very suboptimal as the correct word to pick might not have been the most probable (it might have been the second or third most likely). Therefore, it highly effects the dependence between the predicted words : picking the wrong predicted word will cause the following words to be predicted poorly.

An alternative to greedy search could be beam search as presented by the ACL tutorial. Instead of only picking the token with the highest probability, we pick the K most likely tokens and expand them into K possible word sequences. The best possible sequence is then selected. This method imposes to find the most optimal K value and is much more computationally expensive compared to greedy search. The results are however much better.

2 Question 2

Our translations tend not to end correctly. Indeed, we observe a lot of repetitions for the last few words of some translations. Here is an example :

> I can't help but smoking weed $\longrightarrow$ je ne peux pas empêcher de de *fumer fumer fumer fumer fumer fumer fumer fumer fumer fumer urgence urgence urgence urgence urgence urgence . urgence urgence . urgence urgence* .

This phenomenon is particularly observed for input sentences without any ending punctuation ('.', '!', ...), the model keeps adding the most likely word at the end of the sentence (the last one it has translated) until the end of the sequence. We also observe excessive punctuation repetitions. Here is an example :

> I am a student. $\longrightarrow$ je suis étudiant . . . . . .

Overall, very few sentences present the end-of-sequence (<EOS>) token, marking the end of the translation. The model seems to struggle identifying when to end the translation.

A possible solution to this issus is the local attention model described in [2]. It suggests to focus only on a small, fixed window of the source positions when generating each target word. Instead of attending to all words in the source sentence, it selects a specific window of context around the target word. This approach is more computationally efficient than the global attention model which makes it more adequat for translating longer sequences.

Another approach proposed by [4] is the coverage model. It suggests implementing a coverage set, initially set to zero : $\mathcal{C} = \{0, 0, 0, ...\}$. This set keeps track of which source word has already been translated and prevents the model from translating them again (if a threshold is defined). The coverage set in the end should be : $\mathcal{C} = \{1, 1, 1, ...\}$ representing full coverage (ie. all source words were translated once).

3 Question 3

In Figure 1 we can see a few examples of alignement visualizations :

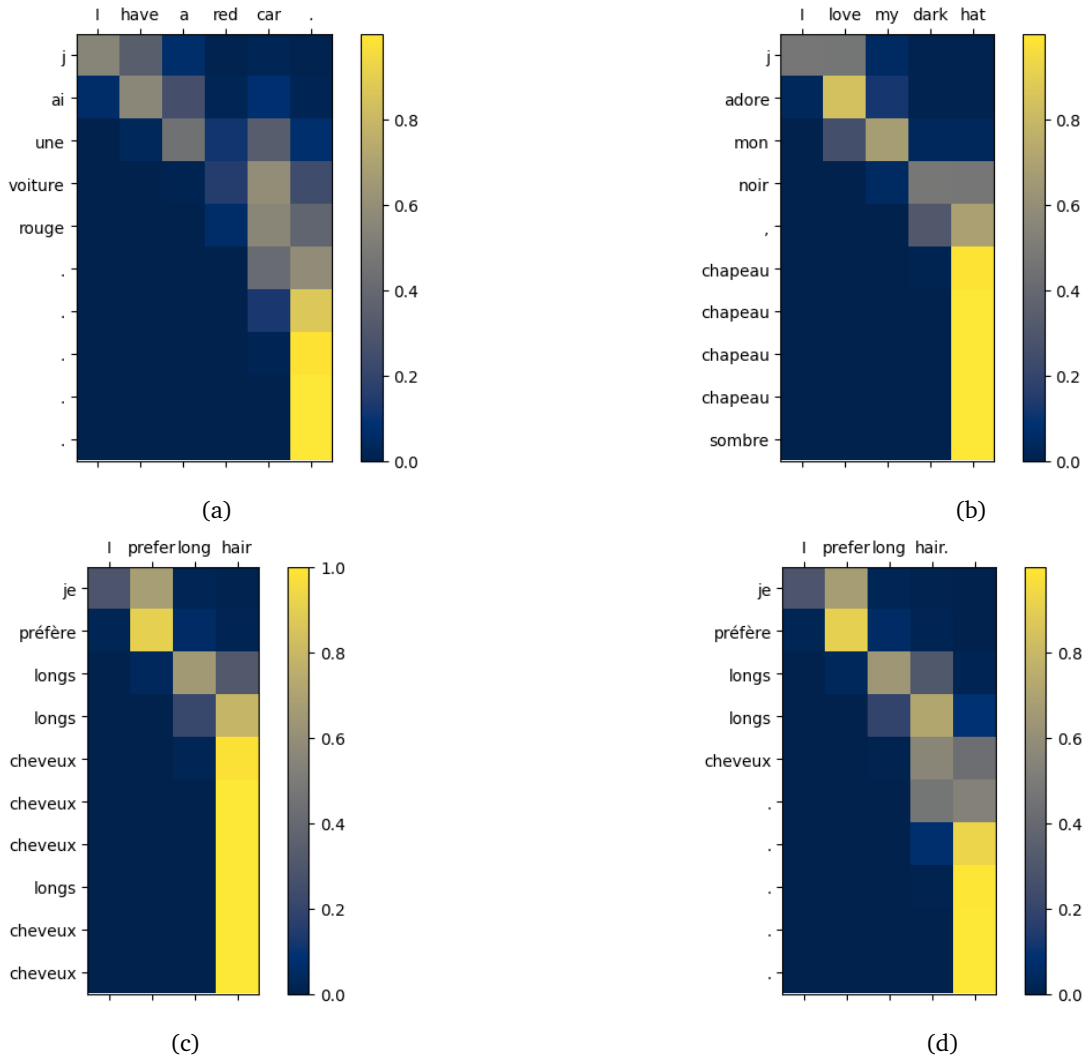


Figure 1: Examples of alignement visualization

The sub-figure (a) shows a successful example of adjective-noun inversion. We notice that the attention for red is higher with ”voiture” that it is for ”rouge” which might suggest the model has managed to capture the link between the color and the object (adjective and noun) therefore translating the sentence correctly.

On the contrary, sub-figure (b) shows an unsuccessful adjective-noun inversion. In this case, the attention for dark with ”chapeau” is very small if not null. Therefore the algorithm translated the sentence literally, word by word without making sense of the dependences and links between words. This might have been caused by an insufficient training of the model.

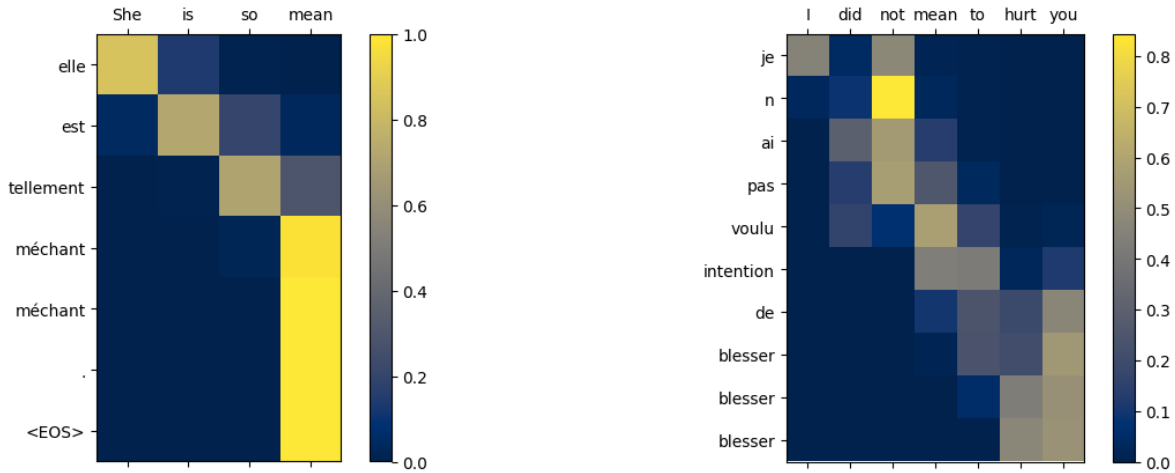
The same thing can be said for sub-figures (c) and (d) in terms of adjective-noun inversion. These two figures also show the difference between punctuated sentences and sentences without puncuation. If both translations are subject to the repetition issue discussed in Question 2, they are different nonetheless. Without punctuation, the translated sentence keeps repeating the last few fords (”cheveux”, ”longs”). On the other hand, the translation of the puctuated sentence repeats the punctuation (’.’).

4 Question 4

Here are the translated sentences followed by their attention weight visualization :

I did not mean to hurt you ⟶ je n ai pas voulu intention de blesser blesser blesser blesser blesser blesser . blesser . blesser . . . . . .

She is so mean ⟶ elle est tellement méchant méchant . <EOS>



(a) (b)

Figure 2: Alignement visualization

Here we notice that the model has correclty captured the difference between the verb 'mean' (=avoir l'intention de, vouloir) and the adjective 'mean' (=méchant(e)). This means the model has been able to include the context of the word 'mean' into its translation.

This example shows how important the context is in translating a specific token. Therefore, to further improve the context inferring of our model, we might want to look at more complex solutions such as BERT described in [1]. This article suggests using a transformer based attention model allowing "bidirectional representations from unlabeled text" that takes into account the context both preceding and following the token that is being translated.

This concept of bi-directional context inferring is also developped in [3] where a method is described for applying it to already existing sequential models such as LSTM or GRU (RNN).

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015.

[3] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018.

[4] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Coverage-based neural machine translation. *CoRR*, abs/1601.04811, 2016.