Gabriel Athènes
gabriel.athenes@polytechnique.edu

Lab session # 2
ALTEGRAD 2021

11/23/21

1 Question 1

The pros of our greedy decoding strategy is that it is very efficient both in terms of memory and computation. Indeed, it picks only one word at the time, the one that has maximum probability given the probability law returned by the decoder at previous time. On the other hand, as the slides point out, this is heavily subobtimal. If a mistake is made at the beginning of a sentence, the mistake is reflected on the rest of the sentence. Another decoding strategy mentioned in the given slides is Beam Search. Beam Search maintains the top k hypothetical translations at every time t, expands all of the translations with all the possible words, and picks again the top k hypothesis at time $t + 1$. This therefore yields to much better quality translations, but is much more expensive too in terms of computations. Looking at the performance of Greedy on test sets, we see that Greedy BLEU performs better than other decoding strategies whereas Greedy NLL performs worse.

2 Question 2

The major problem that we observe in our translations is that words can be translated several times. A way proposed by paper [1] and [3] to tackle this phenomenon is to introduce a coverage set, keeping track of the words we already translated. Indeed, paper [3] proposes to assign to each sentence a list C of integers between 0 and 1 representing which words of the sentence have been translated. For a 4 word sentence, the initial coverage set would be $C = \{0, 0, 0, 0\}$. Once each word has been translated, the coverage list C is equal to $\{1, 1, 1, 1\}$. This coverage set enables us to guide the attention mecanism to make it focus its attention on words that have not been translated yet.

3 Question 3

Let's plot and interpret in the caption the alignment coefficients of two sentences.

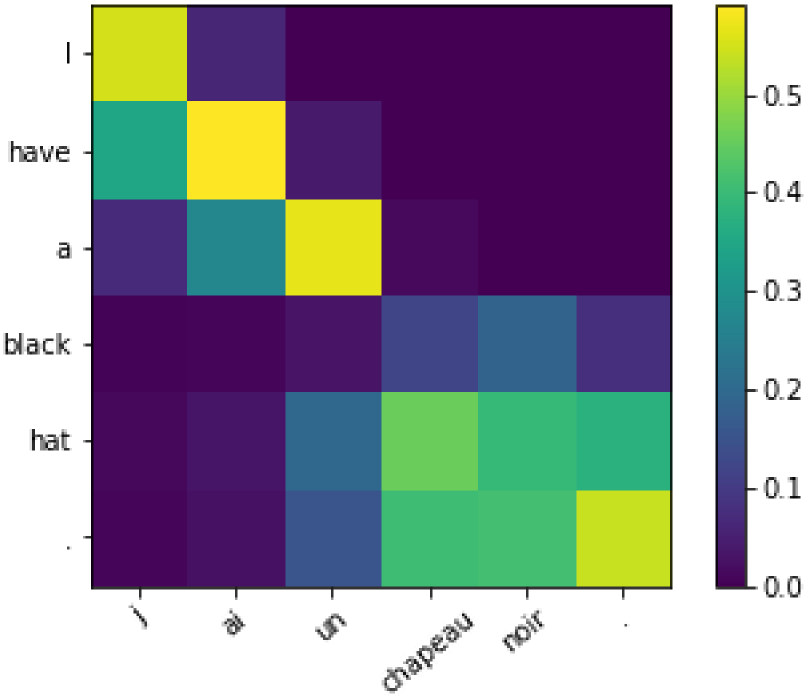

Figure 1: Alignments of « I have a black hat». We can see that the coefficient for « chapeau » is higher for the word « hat », meaning that our model correctly recognized the inversion.

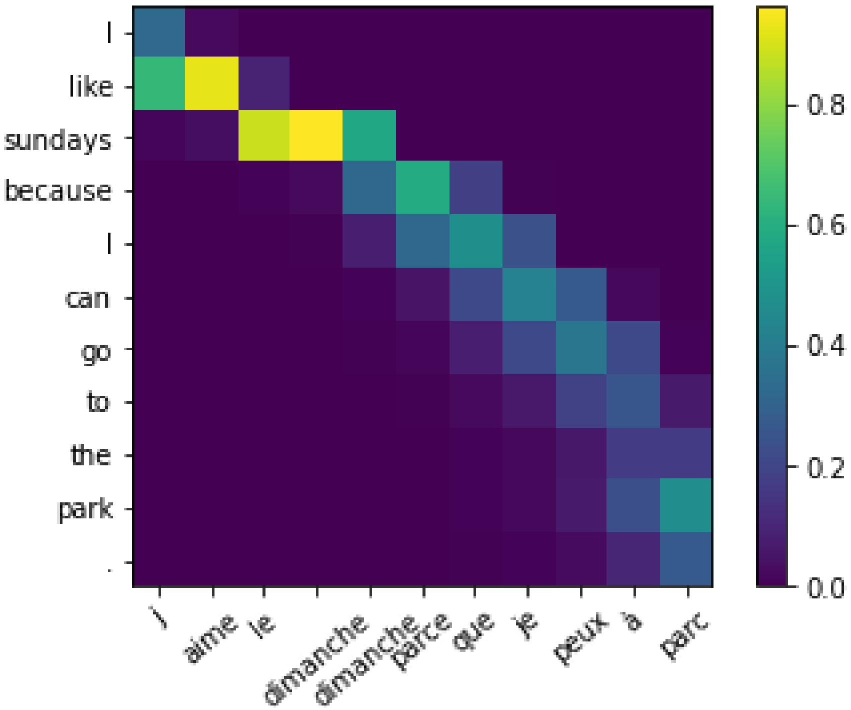

Figure 2: As article [2] explains, in order to model a coverage set for words already translated, we need to take into account the fact that some words are more fertile than others, meaning they translate into more than one word. I therefore gave our model a sentence with many fertile words to analyse the quality of the translation. In the sentence « I like sundays because I can go the park », both « sundays » and « because » lead to a two word translation in french : « J'aime le dimanche parce que je peux aller au parc ». Consequently, we can see that both words « le » and « dimanche » have high coefficients for the word « sundays ». As a result, the model forgets the word « aller » in the translation.

4 Question 4

In the translations of these sentences, we observe that the word «mean » has different translations in french, « méchant » and « intention ». This is due to a property of language called polysemy. One word can have different meanings.

References

[1] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[2] ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. Deep contextualized word representations. arxiv. *arXiv preprint arXiv:1802.05365*, 2018.

[3] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*, 2016.