Mercier Marine
marine.mer98@gmail.com

Lab session # 2
ALTEGRAD 2021

11/23/21

1 Question 1

Greedy decoding strategy is very efficient in terms of computation and memory, the algorithm is really fast, but the quality of the final output sequences may be far from optimal. This strategy select the word with highest probability at each step.

According to [1], some other strategies like the beam search, select multiple words at each step and keep track of multiple sentence hypothesis. This method is compute intensive but give better results.

2 Question 2

The global attention is computationally expensive, some longer sequences can't be translated according to [3]. Indeed all words on the source side are considered for all target words. A local attention mechanism would focus on a subset of the source.

The model we trained tend to repeat words a lot, especially the last words of the sentence. The attention model fails to take benefit from past alignment information and over-translate. A word that has already been translated is less likely to be translated again. In [5], the solution keeps track of which source words have been translated and assigns a lower alignment probability to these words.

The beginning of sentences is well translated, while the end of the translation is messy as the sense is not grasped. We could change the model to a bidirectional Recurrent Neural Network, so that the models doesn't focus only on previous words, but on the entire sentence.

3 Question 3

In this visualization of source and target alignment (cf. Figure 1), we can observe an inversion between the words. "Red car" becomes "voiture rouge", the adjective and noun positions are inverted, therefore the diagonal is not perfect.

4 Question 4

In these sentence, "mean" has a double meaning depending of the rest of the sentence. The model should be able to translate it thanks to the context, but the translation doesn't grasp the sense of the sentence.

Standard language models are unidirectional, therefore some of the context is lost as the model only see previous words. Unlike left-to-right language model pre-training, the solution in [2] is bidirectional and enables the model to train on the left and the right context.

Using deeper neural networks would create a very rich word representation, which would help grasping the subtle difference between these sentences. The representations of words in the solutions described by [4] are a function of all of the internal layers of the biLM.

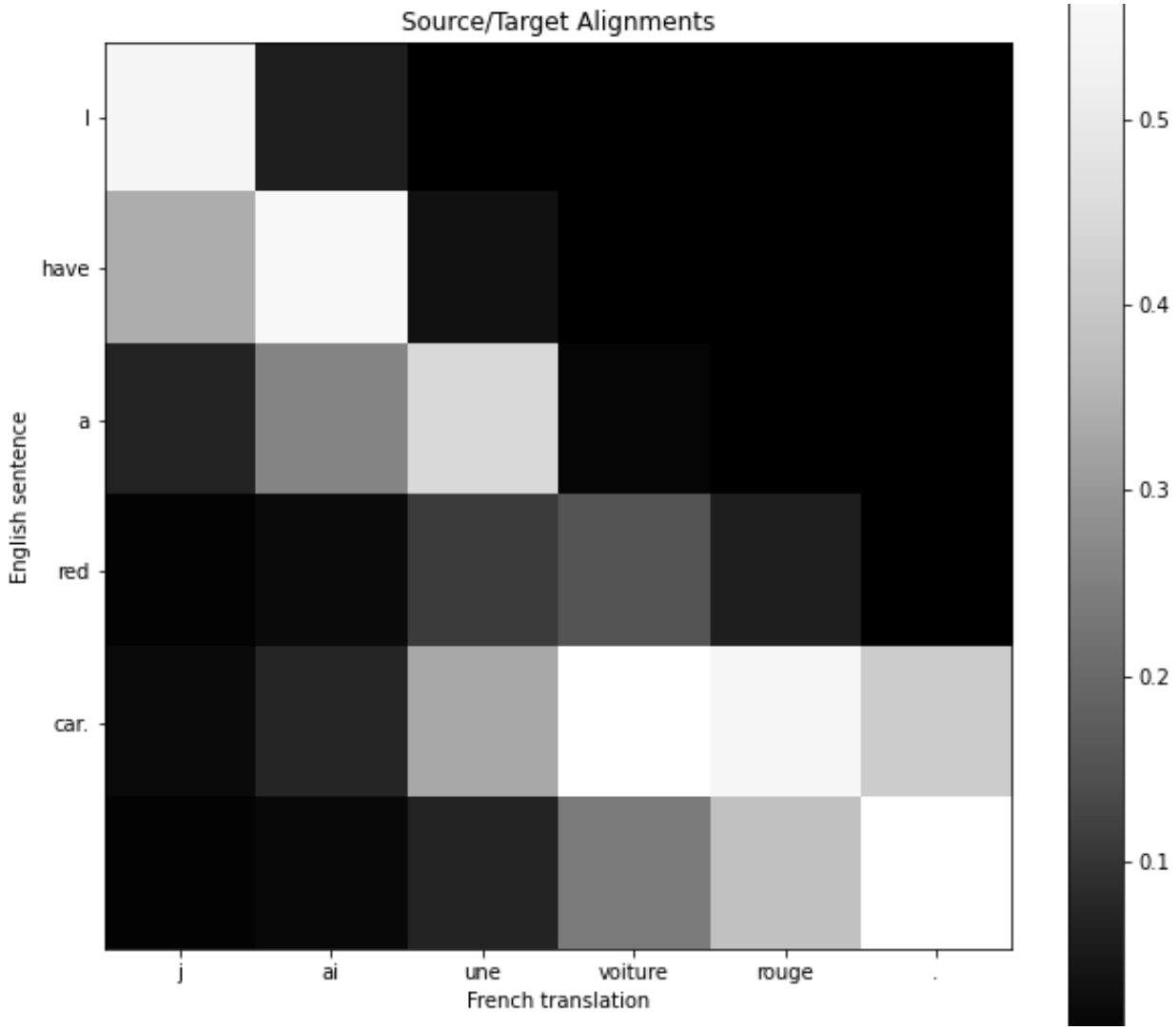

Figure 1: Visualisation of source/target alignment

References

[1] Decoding strategies.

[2] 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.

[3] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[4] 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.

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