Arnaud Massenet
arnaudyoh@gmail.com

Lab session # 4
ALTEGRAD 2019

25/11/19

Question 1

We are currently using a greddy decoding strategy, which has the advantage of being easy to compute. However, as it turns out, if we use the current ”best” word we might end up having no good translations for the words coming afterwards. In essence the ”best” translations might be hiding behind the second or third ”best” word. Instead, we could use beam search, which is less computationally efficient but allows to explore more possible translations by ranking translations and exploring translations using the k ”best” words.

Question 2

The biggest issue in our translations is the inability to translate sentences that are more than around 5 words. This is manifested when the model outputs the same token multiple times. In fact, even if we have correct translations, we still have the issue of repeating tokens. To mitigate this issue [3] proposes to use coverage vectors that keeps in memory which words have been translated and how many times they have been translated. This technique avoids over-translation (when a word is repeated multiple times in the translation) since we can prevent the model from translating words more than n times, where n is some arbitrary number. Another idea comes from [2]. In this paper, the authors present the concept of Input-feeding in contrast with the global attention that we use in our code. This approach allows the model to build its own representation ”coverage” via a deeper architecture.

Question 3

In Fig. 1 we can observe the attention weights of each word in French for each word in English. Note in the top two sentences (*I have a red car*, *My Brown Truck*) that the model is able to properly invert the advectives and nouns. However, there are also examples where it is unable to do so such as in the sentence *Black Cats bring bad luck*. We believe that this is due to non-sufficient training data on such examples.

Question 4

While the sentences are subject to over-translation, the words in the translation are correct for the word ”mean”. This indicated that the models are capable of correctly inferring translation from their context. In the BERT paper [1] the authors present a new architecture based on Transformers which are themselves based on the attention mechanism that we used in this Lab. It is not surprising that this state-of-the-art solution emphasise the importance of context by taking into account the left- and right- context of words in the sentences.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[3] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany, August 2016. Association for Computational Linguistics.

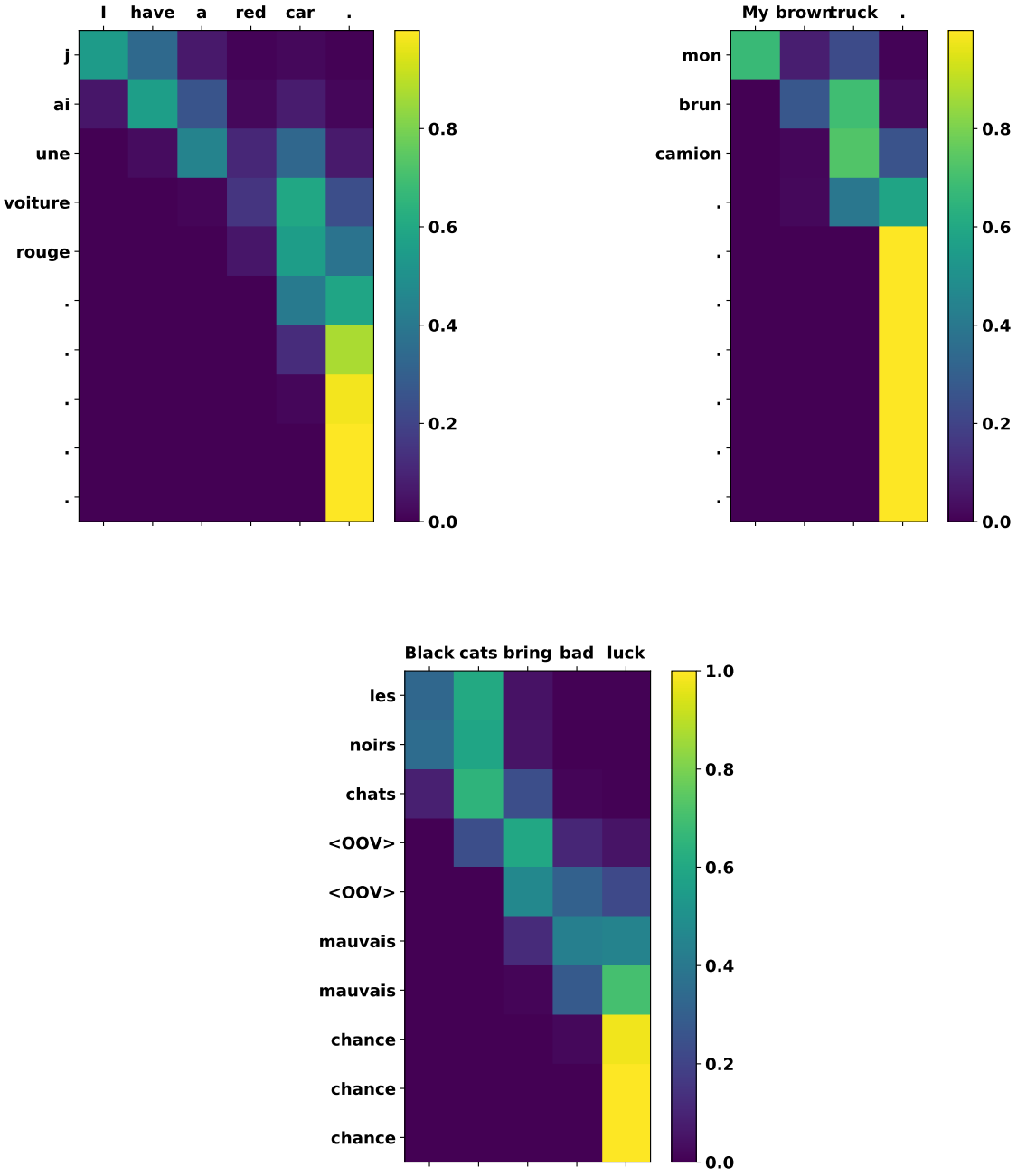

Figure 1: Figure 1: Attention Alignment for the sentences *I have a red car.*, *My Brown Truck.*, *Black Cats bring bad luck*. Note that we limited the lengths of the outputs to make the graphs more readable