Kyllian Asselin de Beauville
kyllian.asselin-de-beauville@polytechnique.edu

Lab session # 1
ALTEGRAD 2023

17/10/23

1 Question 1

In the lab, we employed the greedy decoding strategy, often referred to as the greedy search, for our neural machine translation task. The essence of this approach lies in selecting the word with the highest probability from the model's output at each time step t to be the decoder's output for that specific step, which also serves as the input for the next time step $t + 1$.

Mathematically, for each time step t, the chosen word $\tilde{x}_t$ is:

$$\tilde{x_t} = \arg\max_x P(x|x_{<t}, Y) \tag{1}$$

The primary advantage of the greedy decoding strategy is its computational efficiency. At each time step t, only the top candidate word is considered, making the process swift. However, this efficiency comes at a cost. The strategy is inherently suboptimal because it explores only a single path out of a myriad of potential translation sequences. Moreover, by focusing only on the most probable word at each step, the global structure of the sentence can be overlooked. This can sometimes lead to translations that are grammatically incorrect or with word repetitions.

Among the alternatives presented in this presentation (slides 87-95) from this ACL tutorial, beam search stands out as a particularly notable method. In beam search, instead of limiting the exploration to a single path, the decoder considers the top k probable sequences at each time step, where k is the beam width. This approach strikes a balance between computational efficiency and the quality of the output, offering a higher likelihood of generating syntactically and semantically coherent translations compared to the greedy strategy.

Other decoding methods, such as exhaustive search and ancestral sampling, while theoretically intriguing, may not be practically suitable for our neural machine translation task. Exhaustive search, for instance, considers all possible translation sequences, which becomes computationally infeasible for longer sentences. Ancestral sampling, on the other hand, introduces randomness in the translation process, which may not always yield consistent or accurate translations.

In conclusion, while the greedy decoding strategy has its merits, especially in scenarios with tight computational constraints, other strategies like beam search often offer a more balanced trade-off between translation quality and computational efficiency.

2 Question 2

One of the most apparent issues with our translations is the excessive repetition of words. This repetition problem is not just a superficial flaw but a manifestation of deeper model challenges.

While we used attention mechanisms as proposed by Luong et al. [2], it's evident from our translations that the attention alone might not be sufficient to mitigate this issue. Attention mechanisms are designed to weigh the importance of different parts of the source sentence during translation, allowing the model to focus on relevant portions of the input when producing each word in the output. However, in our case, even with attention, the model appears to overly focus on certain parts, leading to the repeated generation of certain words or phrases.

A potential cause of the excessive repetition might be our greedy decoding strategy. Greedy search selects the most probable word at each step without considering the overall coherence and variability of the entire sentence. This approach can amplify the inherent tendencies of neural models to repeat certain high-probability words or phrases.

To address this, one potential solution could be the introduction of modeling coverage, as described by Tu et al. [4]. Coverage models aim to keep track of which parts of the source sentence have already been translated (”covered”) and discourage the model from repeatedly attending to the same parts. This discouragement is achieved by maintaining a coverage vector. By adding this coverage mechanism, the model becomes more aware of which portions of the source sentence have already been translated, reducing the likelihood of producing repetitive translations.

In conclusion, while attention mechanisms have brought significant improvements to neural machine translation, they are not perfect. Addressing challenges like excessive repetition requires a combination of strategies, including potentially more sophisticated decoding techniques and additional mechanisms like coverage to ensure more accurate translations.

3 Question 3

After updating the code to be able to visualize alignments between source and targets, here are the attention weights for some of our translations.

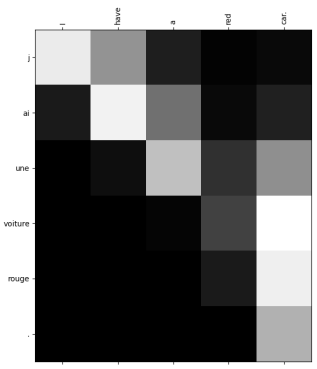

Figure 1: ”I have a red car.”

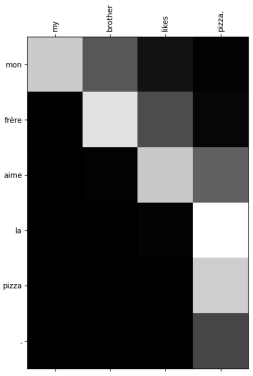

Figure 2: ”my brother likes pizza.”

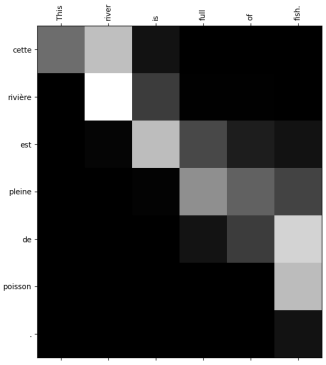

Figure 3: ”This river is full of fish.”

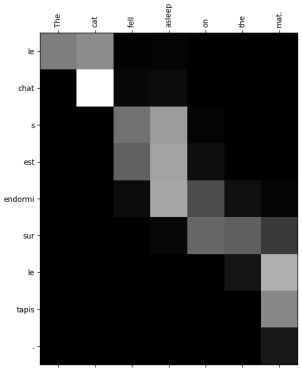

Figure 4: ”The cat fell asleep on the mat.”

Using these visual tools, we can see how the model translates original English sentences to French and identify interesting patterns.

In Figure 1, the model recognizes the different arrangement of adjectives and nouns in English and French. For the translation of ”car” to ”voiture”, it’s aware of this switch. The model also focuses on ”car” and ”red” when translating to ”rouge”, indicating it understands which object is being talked about.

Other images display the model’s skill with typical English-French translation tasks. For instance, in Figure 2, it translates ”pizza” and adds ”la” before it, using ”pizza” as a reference. In Figure 3, the model references ”fish” to translate ”to” as ”de”.

Figure 4 offers more insight. Here, the model translates ”fell asleep” to ”s’est endormi”. Observing how it shifts focus between words gives us clues about its translation process.

4 Question 4

Let’s start by visualizing the translations of these two sentences.

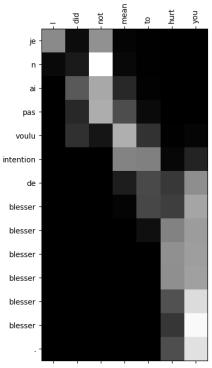

Figure 5: ”I did not mean to hurt you”

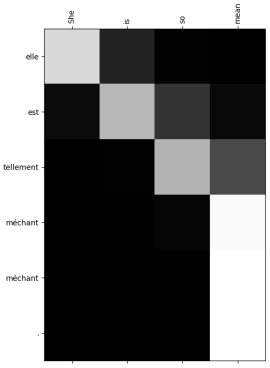

Figure 6: ”She is so mean”

The two sentences, ”I did not mean to hurt you” (Figure 5) and ”She is so mean” (Figure 6), show the polysemous nature of the word ”mean.” In the first, ”mean” is a verb talking about intention. In the second, it’s an adjective describing someone’s nature.

Our model manages to tell the difference between these uses. It knows when ”mean” is about intention and when it’s about nature, even in the translated French version. But when sentences become more complicated, the model might get confused.

One way to help the model is by drawing insights from the BERT paper [1]. BERT stands for ”Bidirectional Encoder Representations from Transformers”. Instead of just reading text from left to right or right to left, BERT reads in both directions. This bidirectional approach is critical for understanding the context. For instance, in the sentence ”I did not mean to hurt you”, BERT would look at both the words before and after ”mean” to figure out that it refers to intention. This gives it a strong advantage in capturing the contextual meaning of words, especially polysemous ones like ”mean”.

In a similar way, the ELMo model [3] provides another avenue for enhanced language comprehension. ELMo stands for ”Embeddings from Language Models”. While ELMo also values context, it achieves this by producing word vectors that consider both past and future words in a sentence. Mathematically, this is represented as:

$$p(t_1, t_2, ..., t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, ..., t_{k-1}) \tag{2}$$

$$p(t_1, t_2, ..., t_N) = \prod_{k=1}^{N} p(t_k | t_{k+1}, t_{k+2}, ..., t_N) \tag{3}$$

In essence, ELMo generates word representations by considering the entire sentence. This methodology results in word vectors that capture the nuanced meanings of words based on their surrounding context, which is especially beneficial for discerning the different meanings of polysemous terms.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[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, 2018.

[4] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation, 2016.