What does Attention in Neural Machine Translation Pay Attention to?

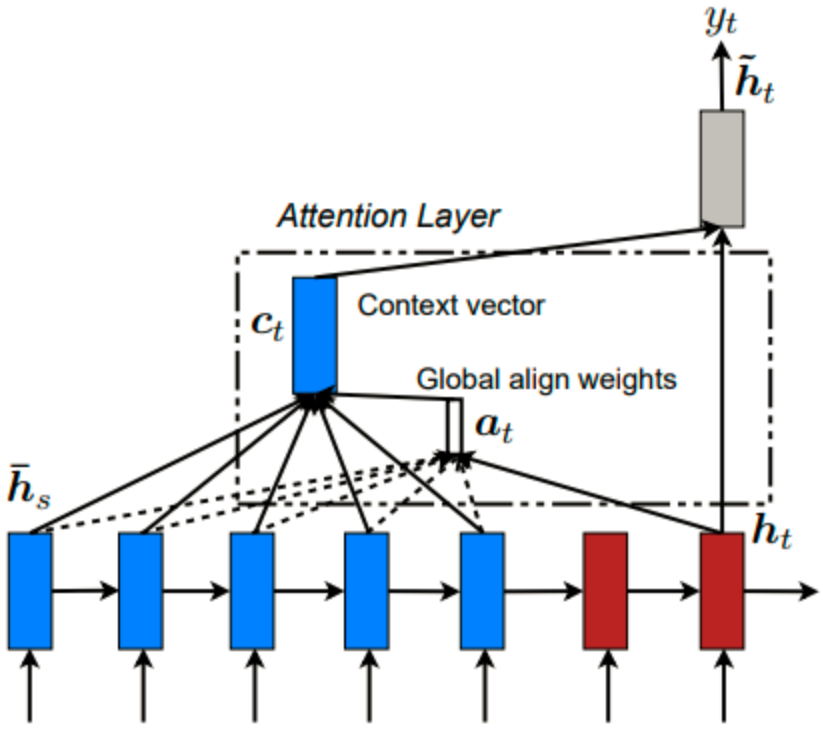
Hamidreza Ghader and Christof Monz
Informatics Institute, University of Amsterdam, The Netherlands

Contribution

This paper analyses the different effections of attention model and aligenment model in NMT, and it shows that attention is different from alignment in some cases and is capturing useful information other than alignments.

Attention Model

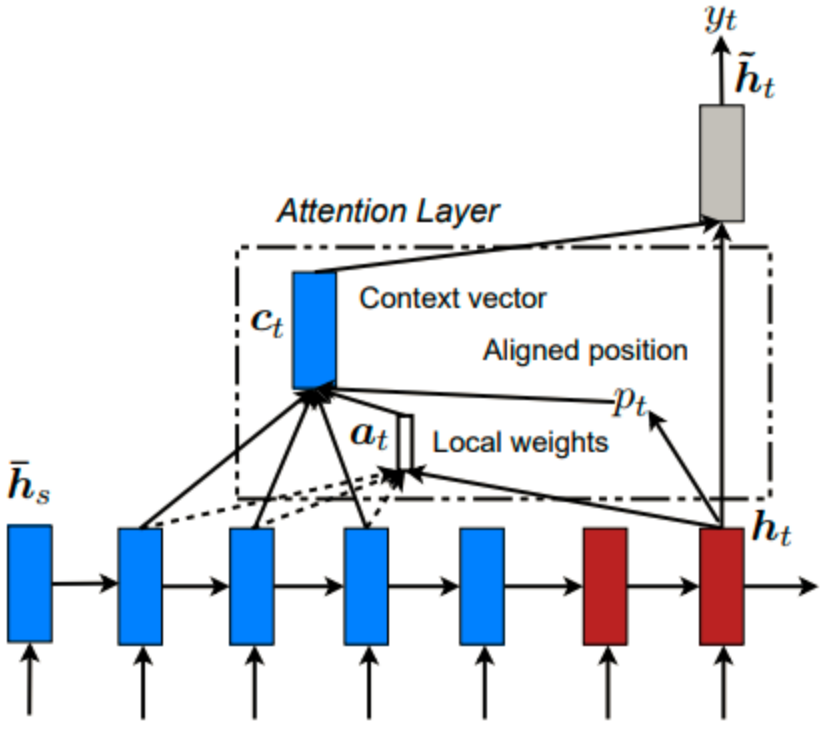
Global attentional model



Attention used in output layer.

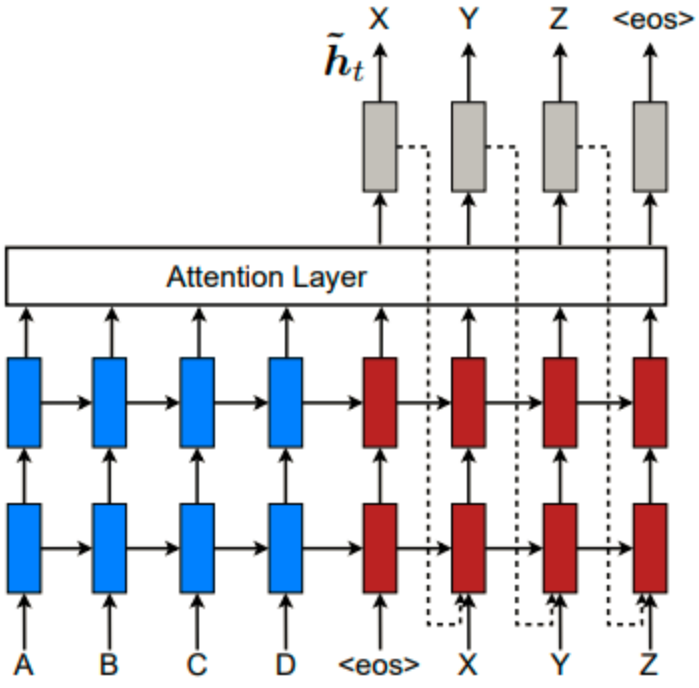
At each time step t, the model infers a alignment weight vector a_t based on the current target state h_t and all source states $\hat{h}_s$.

Local attention model



The model first predicts a single aligned position p_t for the current target word. A window centered around the source position p_t is then used to compute a context-vector c_t, a weighted average of the source hidden states in the window.

Input-feeding attention model



Attentional vectors $\hat{h}_t$ are fed as inputs to the next time steps to inform the model about past alignment decisions.

Measuring Attention-Alignment Accuracy

Using manual alignments provided by RWTH German-English dataset as the hard alignments.

$$Al(x_i, y_t) = \begin{cases} \frac{1}{|A_{y_t}|} & \text{if } x_i \in A_{y_t} \\ 0 & \text{otherwise} \end{cases}$$

Here $A_{yt} is the set of source words aligned to target word y_t and $|A_{yt}|$ is the number of source words in the set.

Attention Loss

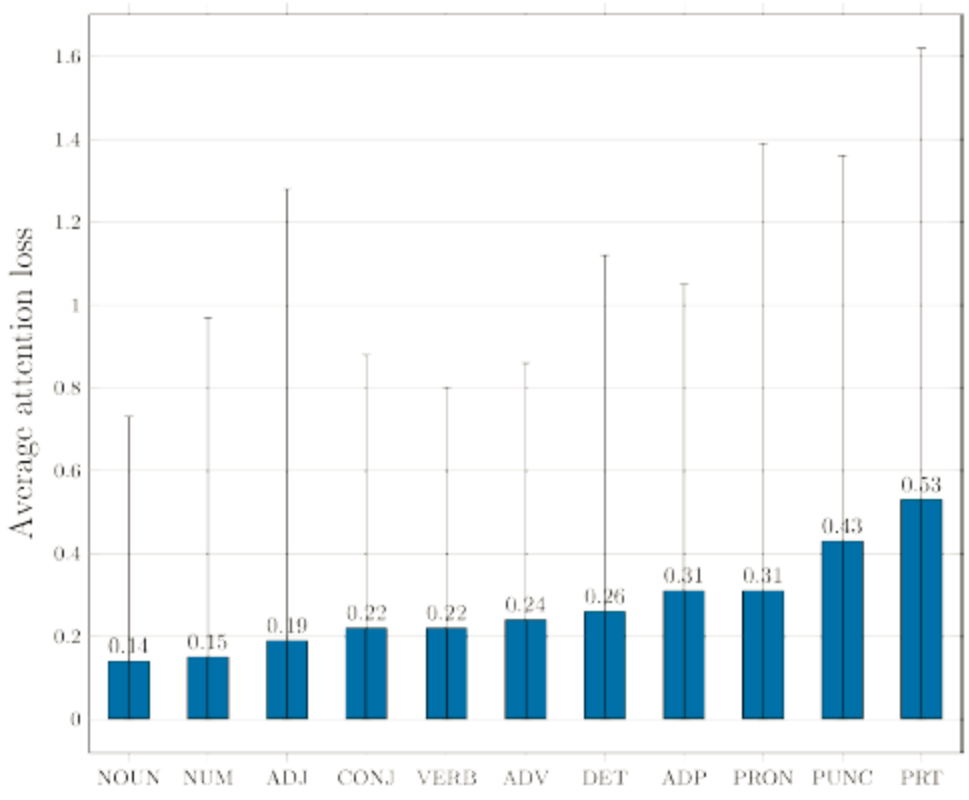
$$L_{At}(y_t) = -\sum_{i=1}^{|x|} Al(x_i, y_t) \log(At(x_i, y_t))$$

Spearman's rank correlation

$$\rho = \frac{\mathrm{Cov}(R_{L_{At}}, R_{L_{WP}})}{\sigma_{R_{L_{At}}} \sigma_{R_{L_{WP}}}}$$

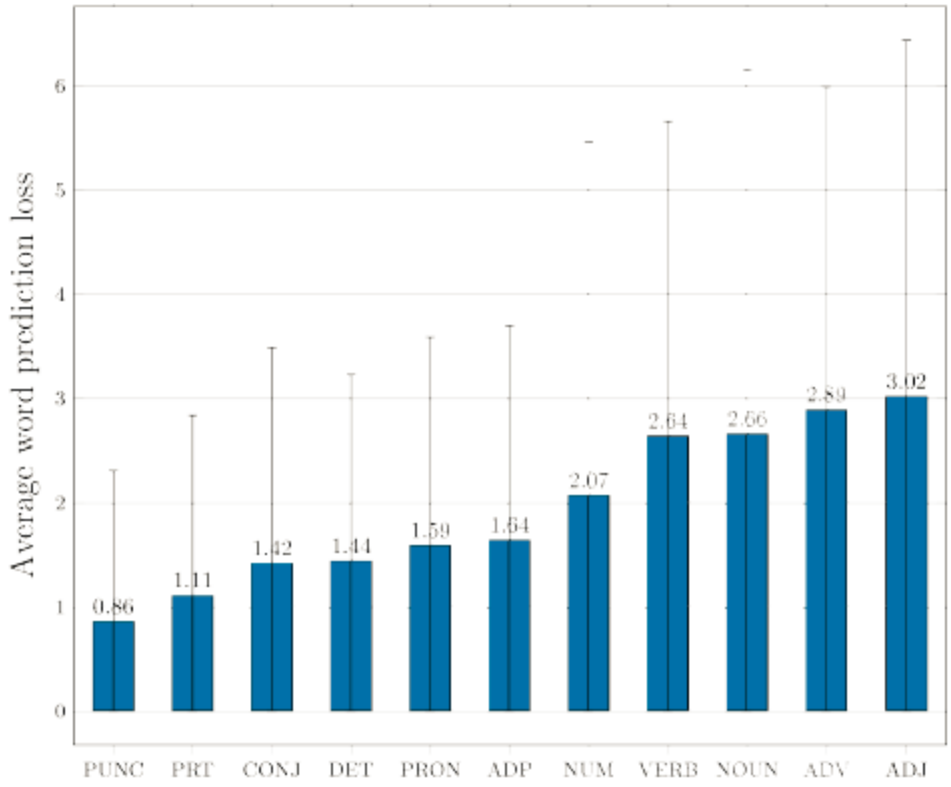
It is used to compute the correlation between attention loss and word prediction loss.
Where $R_{L_{At}}$ and $R_{L_{WP}}$ are the ranks of the attention losses and word prediction losses.

Experiments



(a) Average attention loss based on the POS tags of the target side.

Attention模型在NOUN上起的作用类似于alignment模型



(b) Average word prediction loss based on the POS tags of the target side.

Attention loss低并不能说明word prediction做的好。

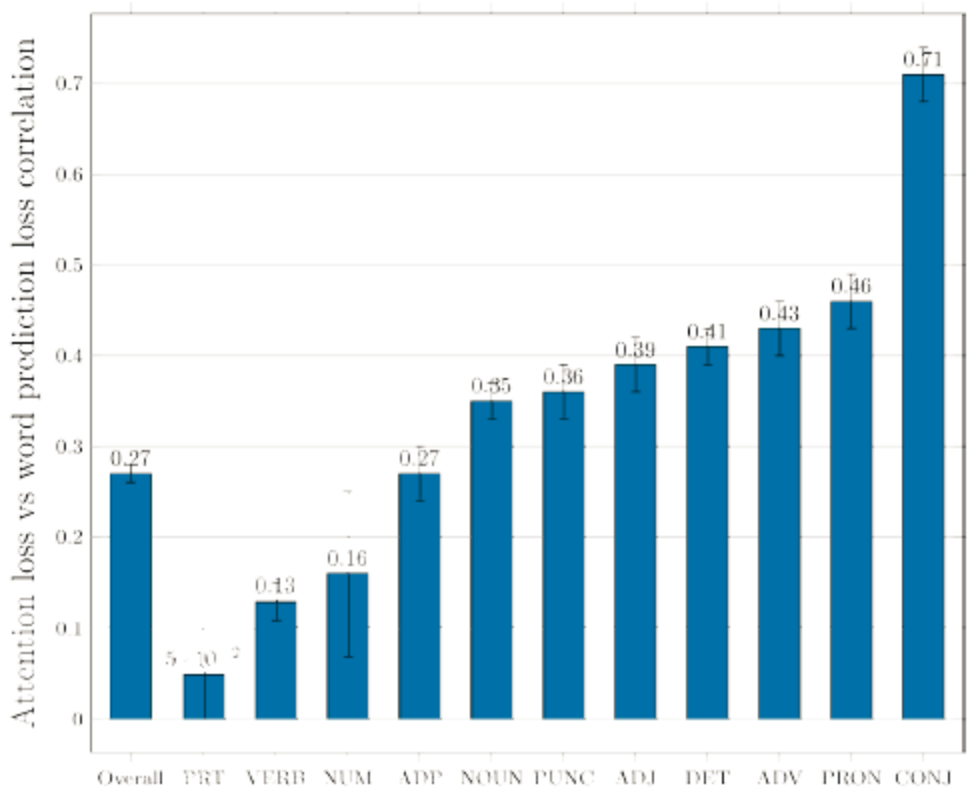


Figure 4: Correlation between word prediction loss and attention loss for the input-feeding model.

对于NOUN等来说，alignment能发挥更好的作用。对于VERB等来说，更加分散的attention能发挥更好的作用。

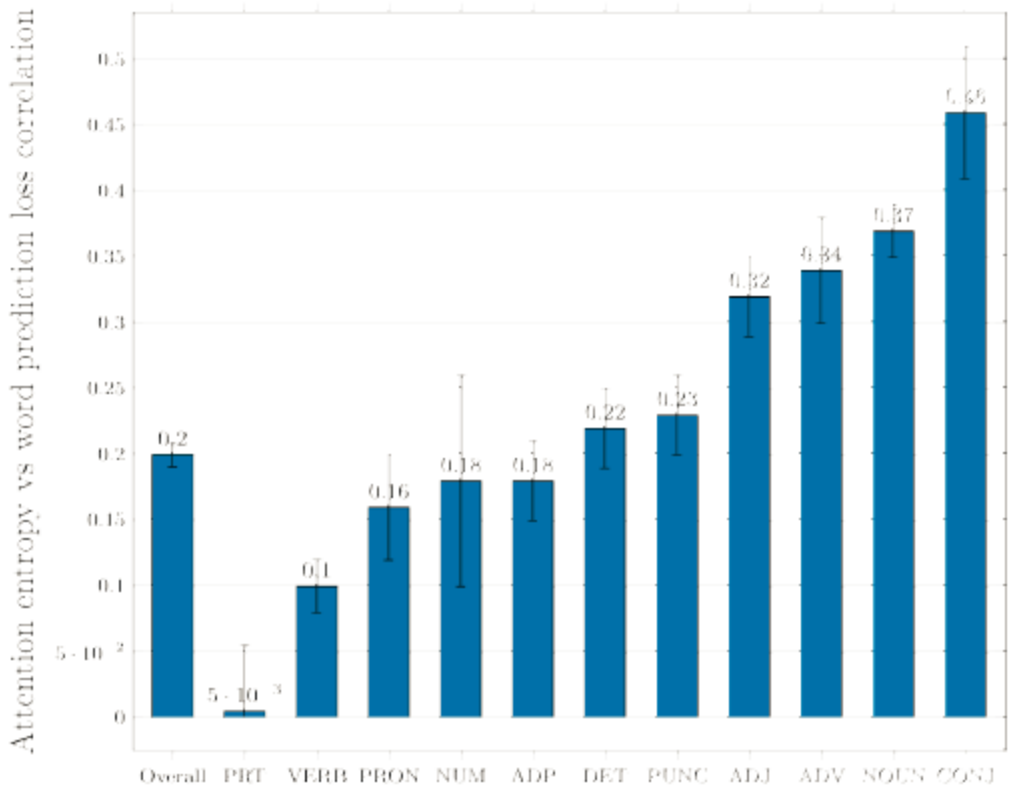


Figure 6: Correlation of attention entropy and word prediction loss for the input-feeding system.

对于VERB等来说，attention越分散，word prediction loss越低，而对于NOUN等来说，attention越分散，word prediction loss越高。

Conclusion

对于名词等的预测来说，我们希望attention越集中越好，但attention model并不能完全实现，所以引入alignment model能更好的预测名词。
而对于动词等来说，我们希望attention更加分散些好，这可以依赖attention model去实现。如果只靠alignment并不能很好实现动词预测。