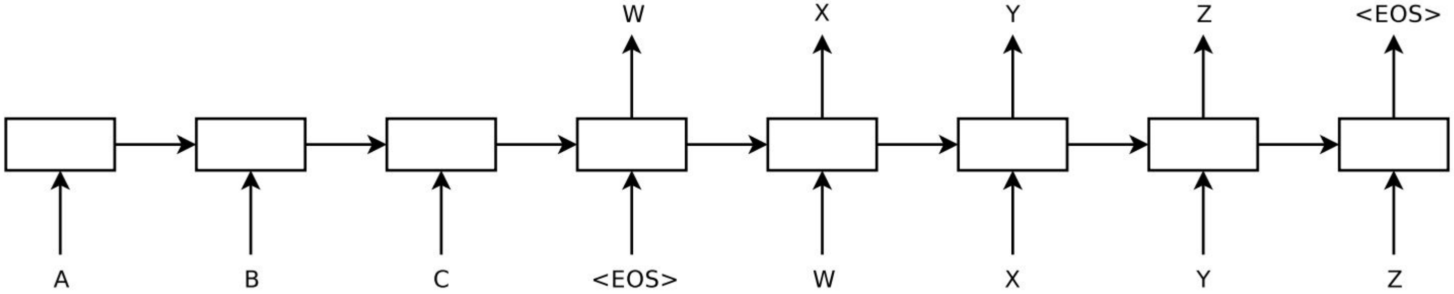
Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong, Hieu Pham, Christopher D. Manning

arxiv.org/abs/1508.04025

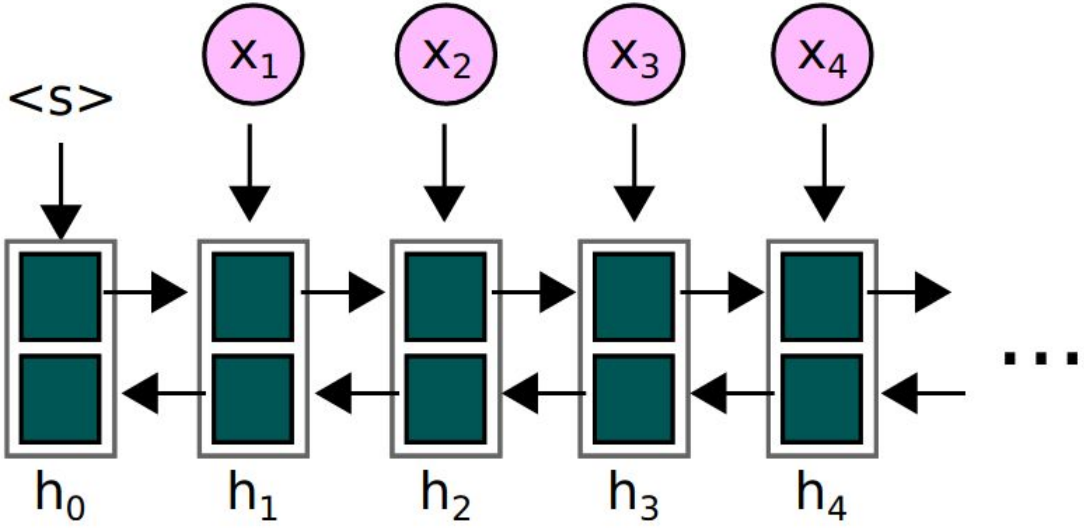
Vanilla Encoder-Decoder



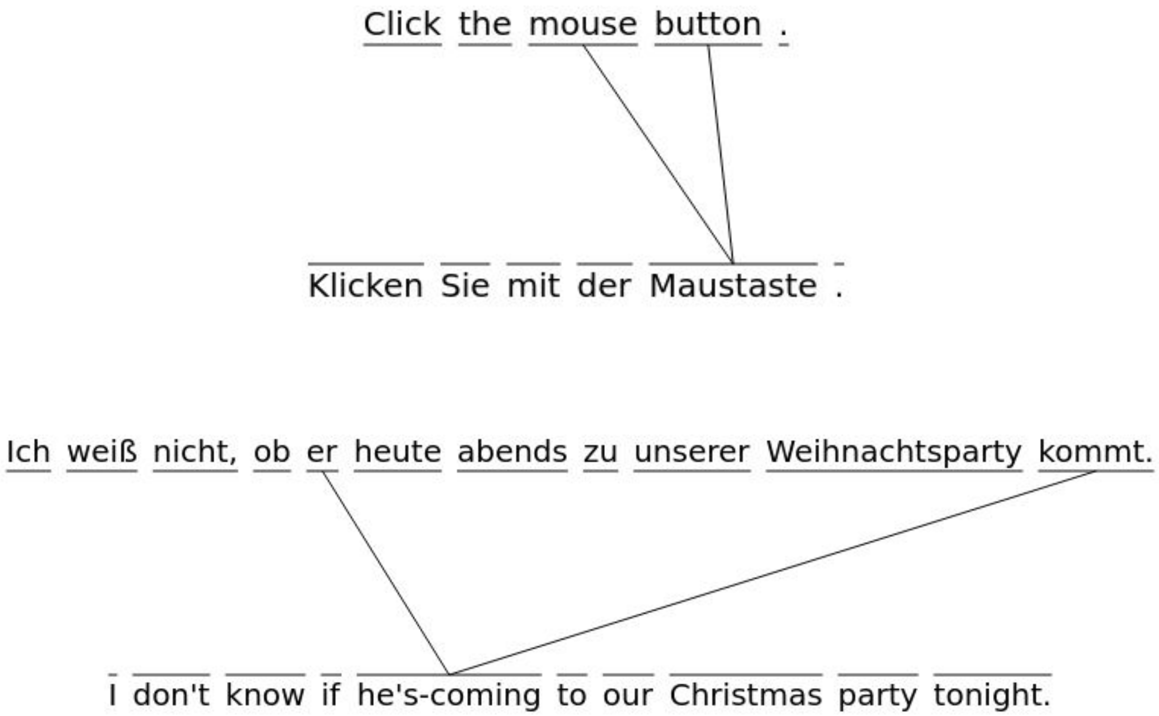
Sentence probability: $\log p(y|x) = \sum_{j=1}^{m} \log p\left(y_j | y_{<j}, \boldsymbol{s}\right)$

Training loss: $J_t = \sum_{(x,y)\in\mathbb{D}} - \log p(y|x)$

BiRNN Encoder-Decoder



Alignment



Problem:
Vanishing gradients, long term dependencies.

Idea:
Tell the network where to look in the input.

Global Attention

Attention relevance between the current decoder state and input state *s*

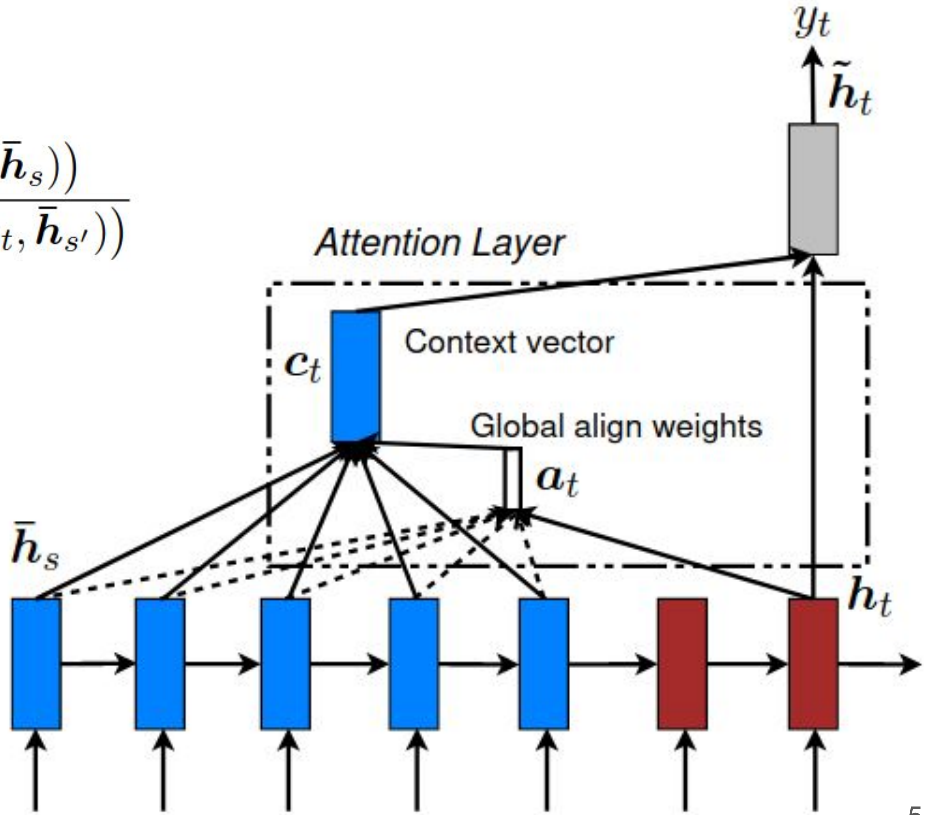
$$\boldsymbol{a}_t(s) = \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)}$$

Context & state concatenation

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W_c}[\boldsymbol{c}_t; \boldsymbol{h}_t])$$

Word output

$$p(y_t|y_{<t}, x) = \text{softmax}(\boldsymbol{W_s}\tilde{\boldsymbol{h}}_t)$$



Computing Score (Globally)

$$\boldsymbol{a}_t(s) = \mathrm{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \frac{\exp\left(\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)}$$

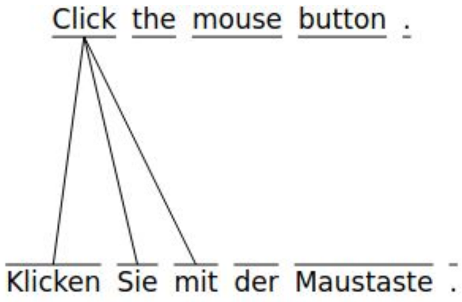
Attention score, not attention!

$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_{\boldsymbol{a}} \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_{\boldsymbol{a}}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$

Attention, not attention score!

$$\boldsymbol{a}_t = \mathrm{softmax}(\boldsymbol{W}_{\boldsymbol{a}} \boldsymbol{h}_t)$$

Local Attention (local-m, local-p)



Diagonal alignment: $p_t = t$

Considered window of input tokens: $[p_t - D, p_t + D]$

Let the network predict the alignment itself:

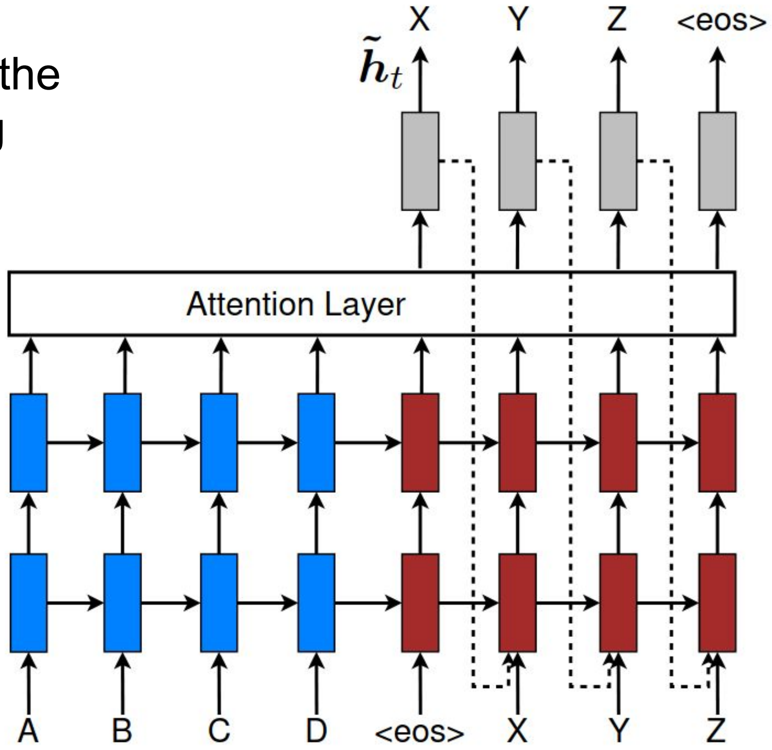
$$p_t = S \cdot \mathrm{sigmoid}(\boldsymbol{v}_p^\top \tanh(\boldsymbol{W_p h}_t)),$$

Force it to focus on tokens around the predicted location:

$$\boldsymbol{a}_t(s) = \mathrm{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right)$$

Input-feeding Approach

Also add the previous hidden state (not the generated word) to the current decoding computation.

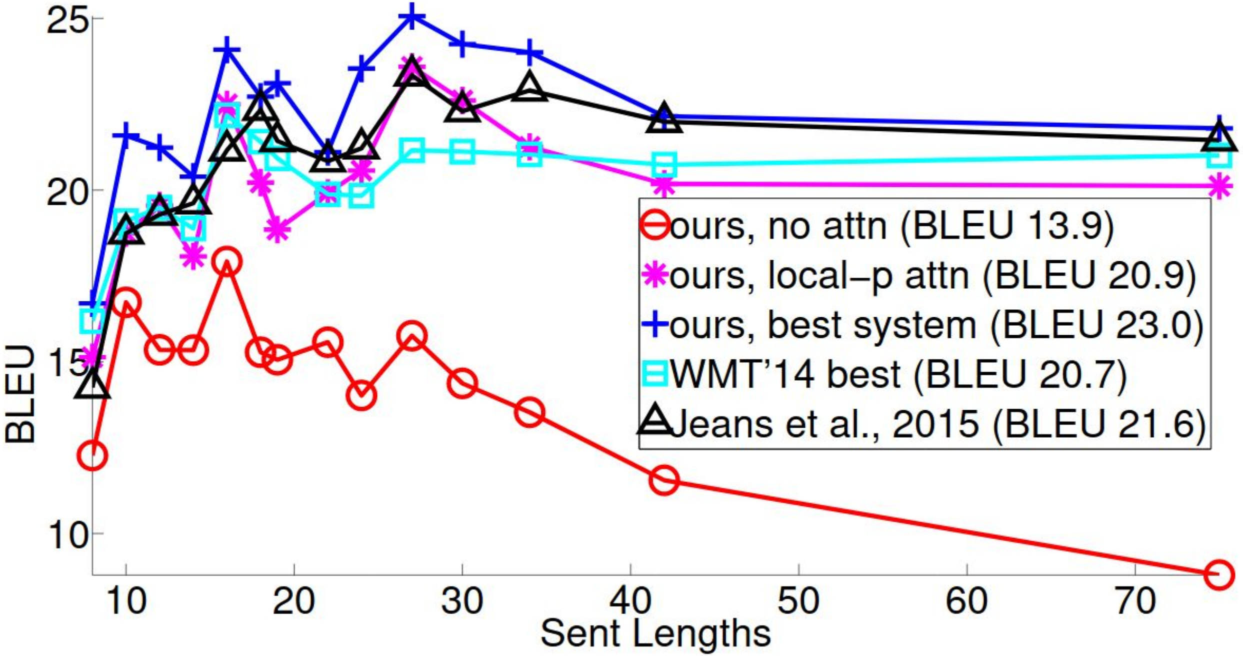


Experiments

System	Ppl	BLEU
Winning WMT'14 system – *phrase-based* + *large LM* (Buck et al., 2014)		20.7
Existing NMT systems		
RNNsearch (Jean et al., 2015)		16.5
RNNsearch + unk replace (Jean et al., 2015)		19.0
RNNsearch + unk replace + large vocab + *ensemble* 8 models (Jean et al., 2015)		**21.6**
Our NMT systems		
Base	10.6	11.3
Base + reverse	9.9	12.6 (+*1.3*)
Base + reverse + dropout	8.1	14.0 (+*1.4*)
Base + reverse + dropout + global attention (*location*)	7.3	16.8 (+*2.8*)
Base + reverse + dropout + global attention (*location*) + feed input	6.4	18.1 (+*1.3*)
Base + reverse + dropout + local-p attention (*general*) + feed input	5.9	19.0 (+*0.9*)
Base + reverse + dropout + local-p attention (*general*) + feed input + unk replace		20.9 (+*1.9*)
Ensemble 8 models + unk replace		**23.0** (+*2.1*)

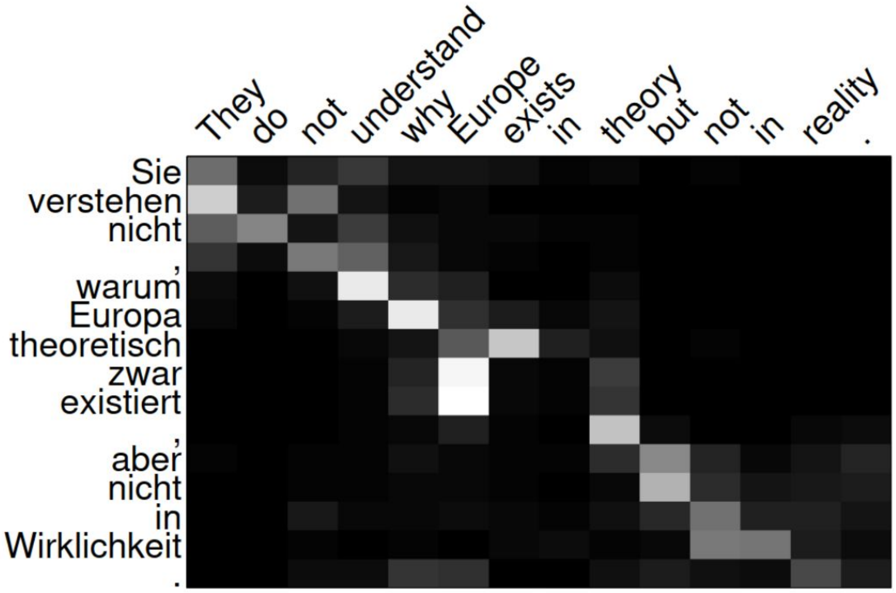
Sentence Length

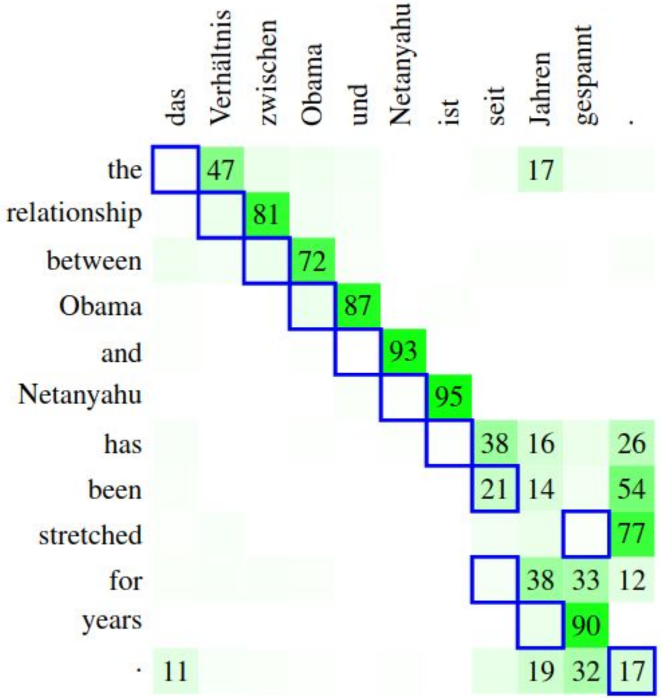
Attention saves long sentences, especially compared to vanilla RNN.



Attention Note #1

Why is alignment off by one?

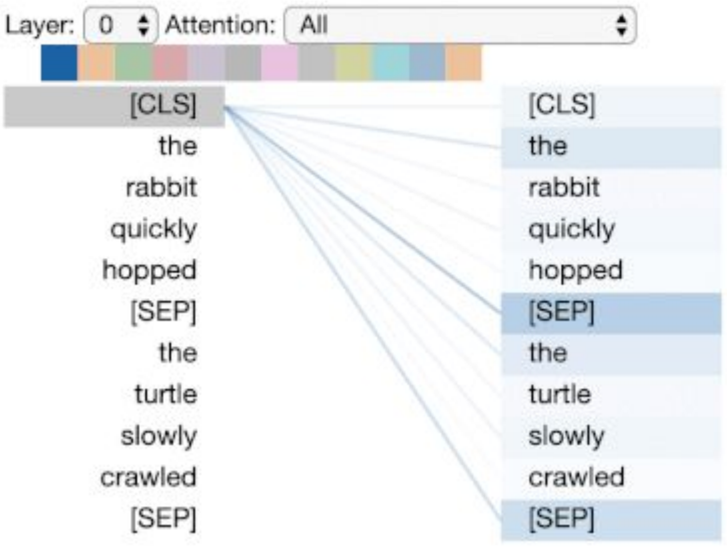




Attention Note #2

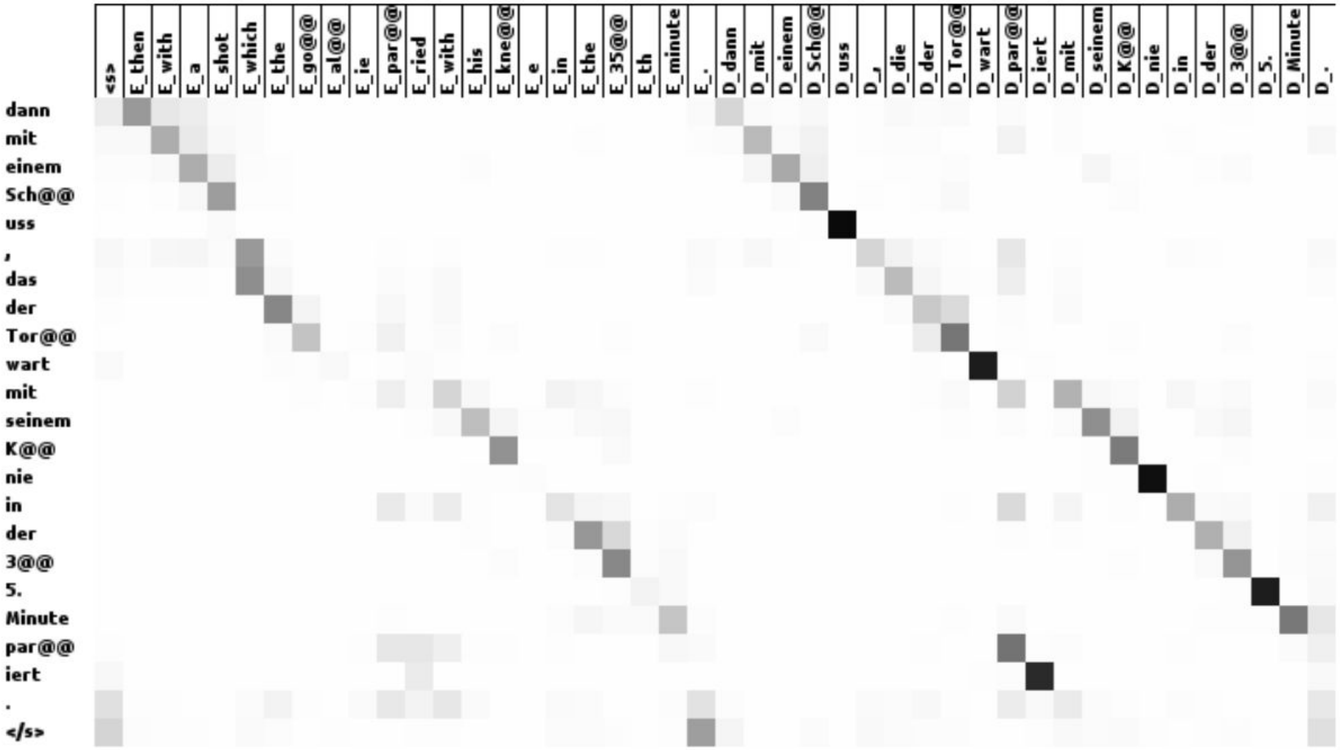
Attention heads encode linguistic structures.

Heads from Transformer self-attention resemble linguistic dependency.



Attention Note #3

Append less performant MT output as an input to attention based MT



Summary

- vanilla RNN suffers from vanishing gradients

- attention brings input closer to the output

- global attention - whole sentence
- local attention - window of tokens

Resources

- Effective Approaches to Attention-based Neural Machine Translation, Minh-Thang Luong, Hieu Pham, Christopher D. Manning
  https://arxiv.org/abs/1508.04025
- Class on Statistical Machine Translation
  Ondřej Bojar
  http://ufal.mff.cuni.cz/courses/npfl087
- Enabling Outbound Machine Translation
  My bachelor thesis
  https://dspace.cuni.cz/bitstream/handle/20.500.11956/119400/130284419.pdf
- Six Challenges for Neural Machine Translation
  Philipp Koehn, Rebecca Knowles
  https://arxiv.org/abs/1706.03872
- A Multiscale Visualization of Attention in the Transformer Model
  Jesse Vig
  https://github.com/jessevig/bertviz
- Slow Align
  https://vilda.net/s/slowalign/