Self-Attentive Residual Decoder for Neural Machine Translation

Lesly Miculicich*†, Nikolaos Pappas*, Dhananjay Ram*†, Andrei Popescu-Belis‡

*Idiap Research Institute, †EPFL, ‡HEIG-VD / HES-SO

Motivation

Limitations of the RNN-based decoder for NMT

- The RNN's internal memory is shared across words and is prone to a recency bias.
- Does not fully capture the structure of language.

Proposed approach:

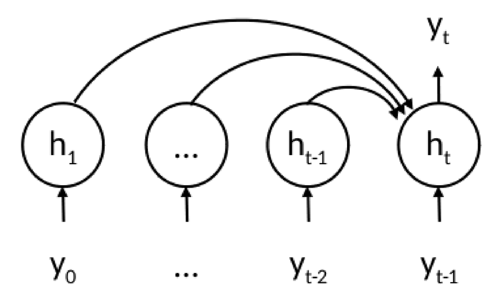
- Enhance the RNN memory with direct and selective access to past.
- The residual connections facilitate the flow of information.
- The self-attention allows selective use of previously predicted words.

Other Self-Attentive Networks

Memory RNN

RNN with memory cells of previous representations

[Cheng et al., EMNLP 2016]

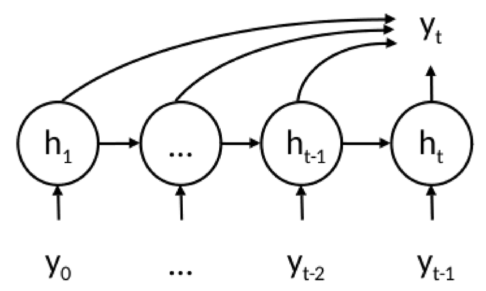


$$h_t = f(\tilde{h}_t, y_{t-1}, c_t)$$
$$\tilde{h}_t = \sum_{i=1}^{t-1} \alpha_i^t h_i$$
$$\alpha_i^t = \text{attention}(\tilde{h}_{t-1}, h_i, y_{t-1})$$

Self-Attentive RNN

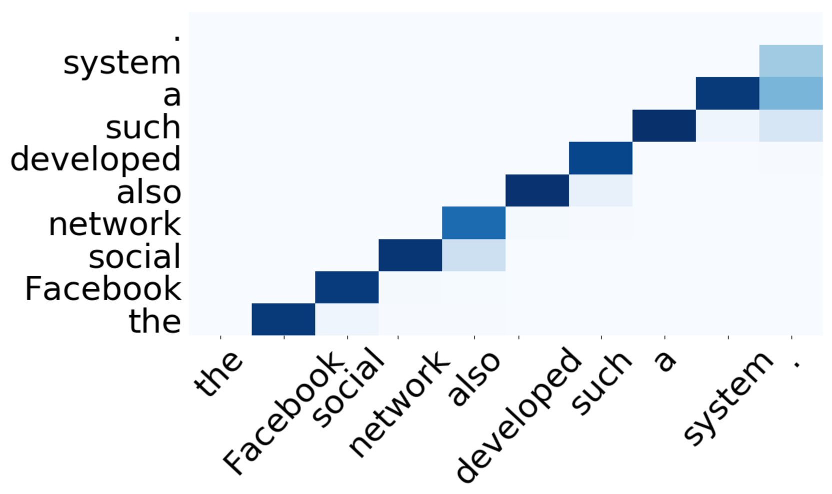
RNN with a summary vector from past predictions

[Daniluk et al., ICLR 2016]

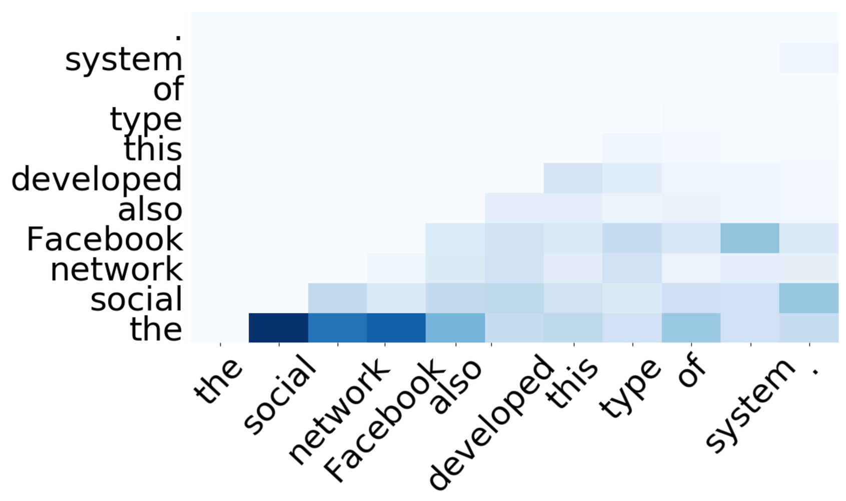


$$p(y_t|y_1, ..., y_{t-1}, c_t) \approx g(h_t, y_{t-1}, c_t, \tilde{h}_t)$$
$$\tilde{h}_t = \sum_{i=1}^{t-1} \alpha_i^t h_i$$
$$\alpha_i^t = \text{attention}(h_i, h_t)$$

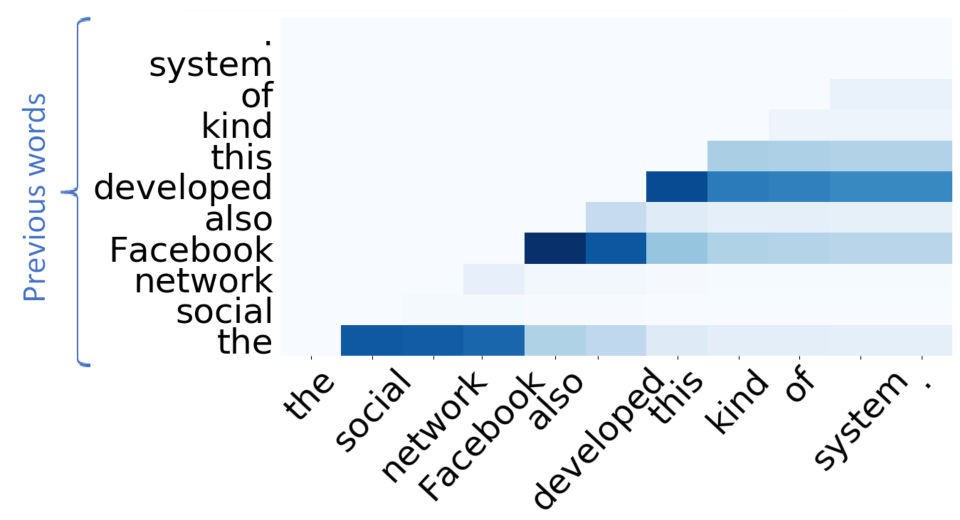
Self-attention Matrices:



Memory RNN: attention mainly to the previous word



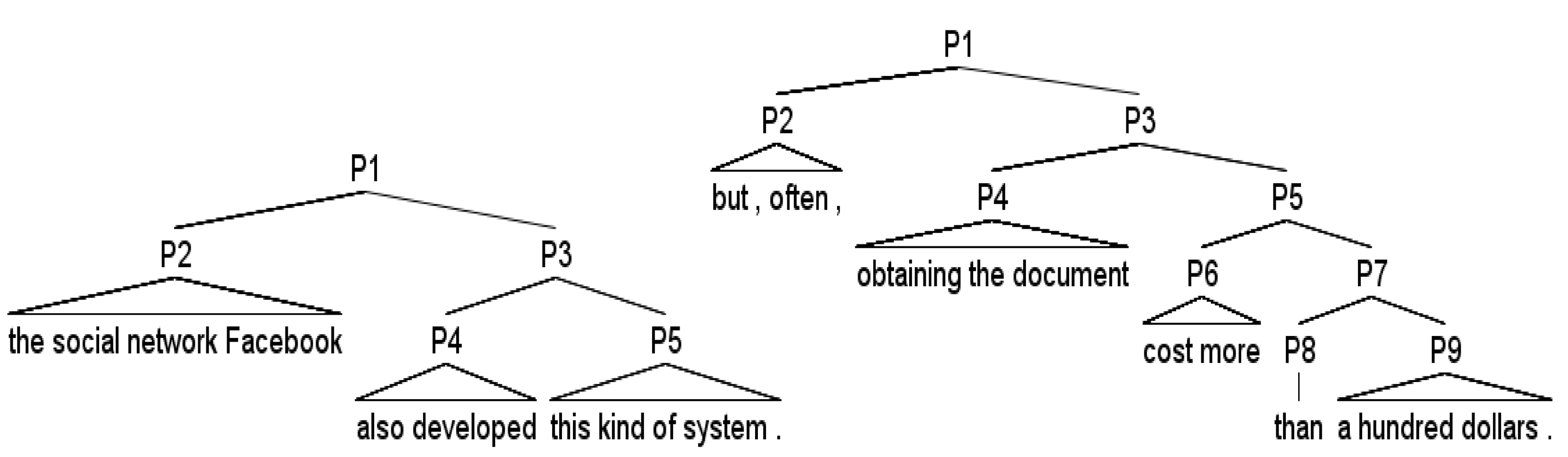
Self-attentive RNN: attention to first words of the sentence



Self-attentive residual connections: attention to different marker words

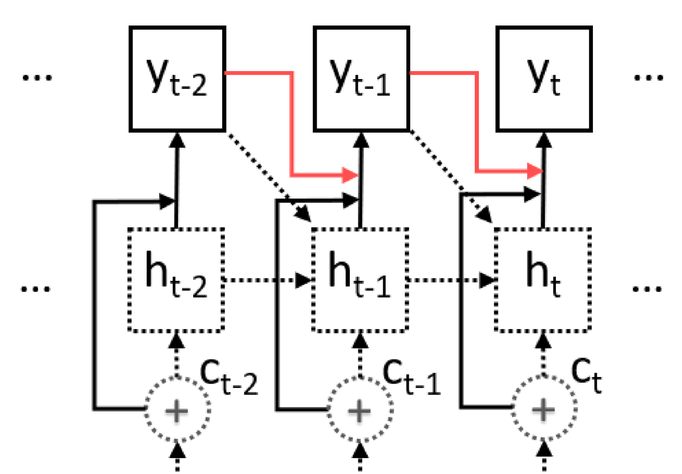
- Formation of "phrases" when grouping words by their focus of attention.

Hypothesized Syntactic Structures:



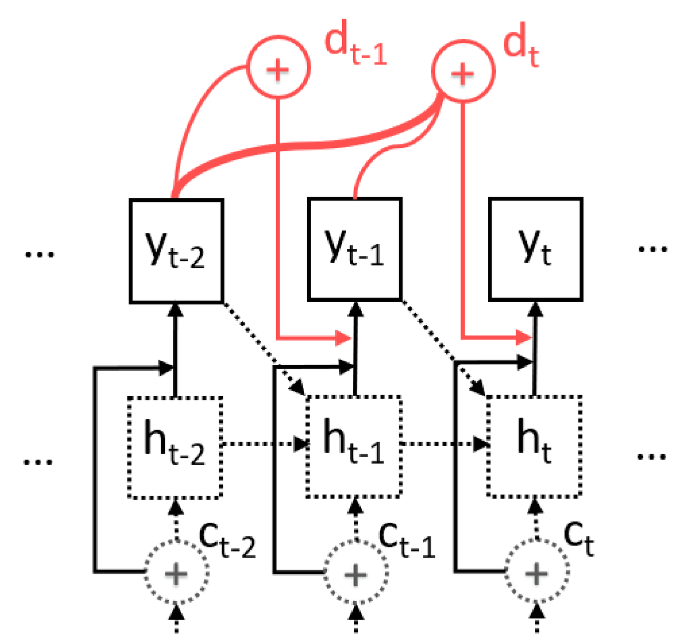
- The trees are obtained from the attention weights of the self-attentive residual connections through a binary tree parser algorithm.

Self-Attentive Residual Decoder



Baseline NMT decoder

$$p(y_t|y_1, ..., y_{t-1}, c_t) \approx g(h_t, c_t, y_{t-1})$$
$$h_t = f(h_{t-1}, y_{t-1})$$



Self-attentive residual decoder

$$p(y_t|y_1, ..., y_{t-1}, c_t) \approx g(h_t, c_t, d_t)$$
$$h_t = f(h_{t-1}, y_{t-1})$$
$$d_t = f_a(y_1, ..., y_{t-1})$$

- The baseline NMT decoder uses a residual connection to the previously predicted word y_{t-1}
- We propose to use residual connections from all previously translated words $y_1, ..., y_{t-1}$ with a summary vector d_t.

Mean Residual

$$d_t^{avg} = \frac{1}{t-1} \sum_{i=1}^{t-1} y_i$$

Self-Attentive Residual

$$d_t^{cavg} = \sum_{i=1}^{t-1} \alpha_i^t y_i$$
$$\alpha_i^t = \text{attention}(y_1, .., y_{t-1})$$

Experimental Setup

Datasets: En-ZH UN Corpus 0.5M, Es-En WMT 2.1M, En-De WMT 4.5M

Architecture: Attention-based NMT with GRUs of dimension 1024, 500 for word embeddings, and vocabulary of 50K.

Results

	$\lvert\Theta\rvert$	BLEU		
Models		**En–Zh**	**Es–En**	**En–De**
SMT baseline	–	21.6	25.2	23.2
NMT transformer (comparable model)	109.0M	22.0	25.9	24.1
NMT baseline	108.7M	22.6	25.4	24.8
+ Memory RNN	109.7M	22.5	25.5	24.9
+ Self-attentive RNN	110.2M	22.0	25.1	24.3
+ Mean residual connections	108.7M	23.6	25.7	24.9
+ Self-attentive residual connections	108.9M	**24.0**	**26.3**	**25.5**

BLEU on *tokenized* text. $|\Theta|$ is the number of parameters.

- Self-attentive residual connections outperform other models, while using fewer parameters than other self-attentive methods.

Code at: https://github.com/idiap/Attentive_Residual_Connections_NMT

Conclusion

- We proposed self-attentive residual learning framework.
- Improvements over a standard baseline, and two variants of self-attention.
- Analysis of the attention shows syntactic-like structures.
- It can be applied to other tasks based on RNNs.

Acknowledgements

Supported by the European Union Horizon 2020 SUMMA project (grant 688139, www.summa-project.eu).