On The Alignment Problem In Multi-Head Attention-Based Neural Machine Translation

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
<surname>@i6.informatik.rwth-aachen.de

*in WMT 2018

Alignment-based NMT

- Using explicit hard alignments to help translation.
- Useful when customer wants to enforce specific translation of certain words.
- Two steps: alignment generation and word generation.

$$p(e_1^I|f_1^J) = \sum_{b_1^I} p(e_1^I, b_1^I|f_1^J) \qquad (1)$$

$$\approx \max_{b_1^I} \prod_{i=1}^{I} \underbrace{p(e_i|b_i, b_1^{i-1}, e_1^{i-1}, f_1^J)}_{\text{lexical model}} \cdot \underbrace{p(b_i|b_1^{i-1}, e_1^{i-1}, f_1^J)}_{\text{alignment model}}.$$

*: Alkhouli et al., Alignment-Based Neural Machine Translation, in WMT 2016.

Self-Attentive Alignment Model

- Employ Transformer as the alignment model to predict source positions.
  - The output is a probability distribution over possible source jumps:

$$\Delta_i = b_i - b_{i-1}$$

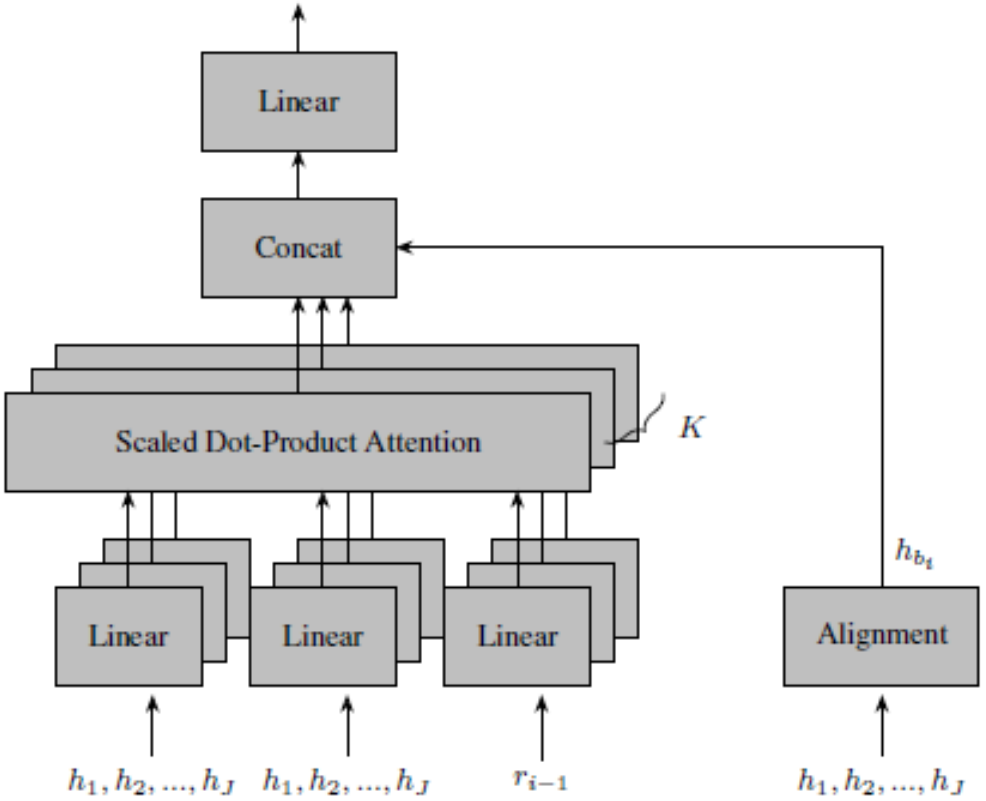
  - Use a single-head hard attention to replace multi-head source-to-target attention.

$$\alpha(j|b_{i-1}) = \begin{cases} 1, & \text{if} \quad j = b_{i-1} \\ 0, & \text{otherwise.} \end{cases}$$

defined for the source positions $j, b_{i-1} \quad \in$

Transformer-Based Lexical Model

- Add an additional alignment head to help generate words.



Experiments

#	System	Layer size	WMT En→Ro newstest2016 PPL	WMT En→Ro newstest2016 BLEU [%]	WMT En→Ro newstest2016 TER [%]	BOLT Zh→En test PPL	BOLT Zh→En test BLEU [%]	BOLT Zh→En test TER [%]
baselines								
1	Attention baseline	1000	10.2	24.7	58.9	8.0	20.0	65.6
2	Transformer baseline	2048	6.2	27.9	54.6	6.0	22.5	62.1
3	(Alkhouli and Ney, 2017)	200	-	24.8	58.1	-	-	-
this work								
4	RNN Attention align.-biased	1000	7.2	26.4	56.1	5.6	19.6	62.3
5	Align.-assisted Transformer	2048	**5.0**	**28.1**	**54.3**	**4.7**	**22.7**	**61.8**

Table 2: Translation results for the WMT 2016 English→Romanian task and the BOLT Chinese→English task. We include the lexical model perplexities.

Dictionary-guided NMT

- More accurate alignment.

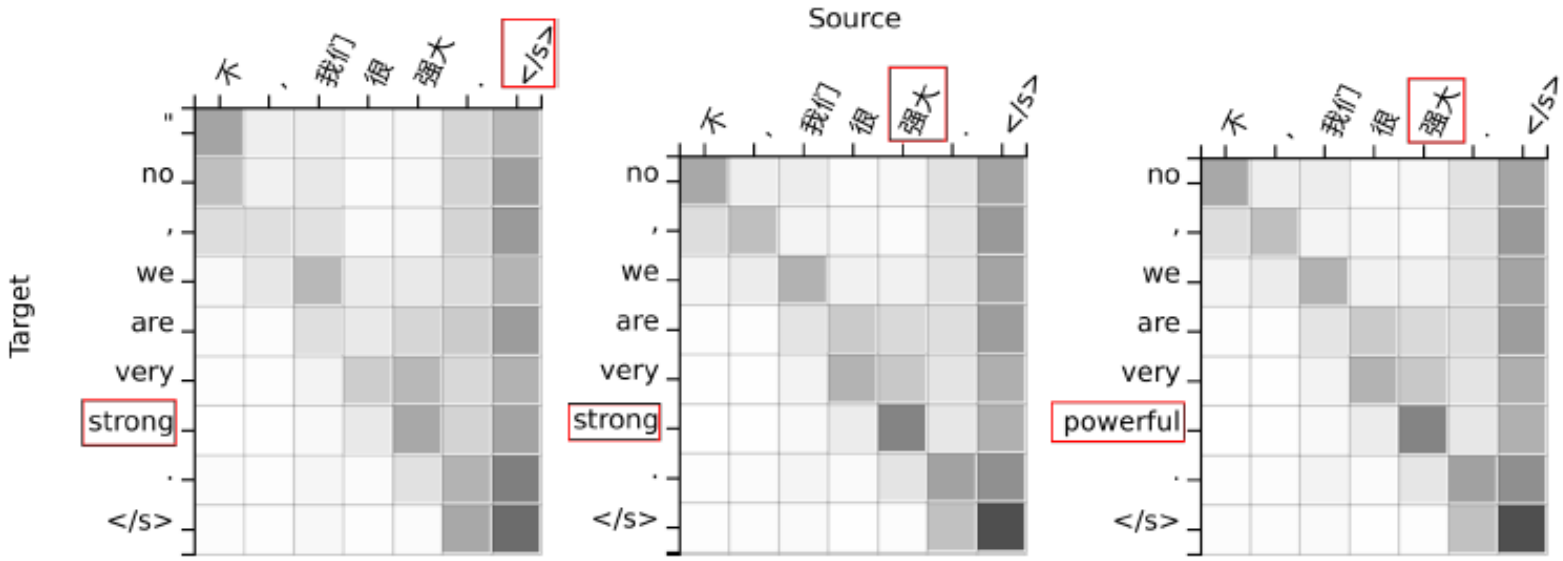


Figure 1: An example from the Chinese→English system. The figures illustrate the accumulated attention weights of the baseline transformer model (left), the alignment-assisted transformer model (middle), and the alignment-assisted model guided by a dictionary entry. We simulate a scenario where the user wants to translate the Chinese word "强大" to "powerful". Both the baseline and alignment-assisted

Inspiration

- Incorporate other source of attention heads, e.g., CNN, RNN, or linguistic features.