Effective Approaches to Attention-based Neural Machine Translation

18/11/10

김보섭

Agenda

1. Introduction
2. Neural Machine Translation
3. Attention-based Model
   - Global Attention
   - Local Attention
   - Input-feeding Approach
4. Experiments
5. Analysis
6. Conclusion

Introduction

본 논문에서는 Neural Machine Translation에서 활용할 수 있는 단순하면서도 매우 효과적인 두 가지의 attentional mechanism을 제안함

- Global approach : all source words are attended

  → Global Attention

- Local approach : only a subset of source words are considered

  → Local Attention

Neural Machine Translation

Neural Machine Translation은 neural network가 source sentence가 주어졌을 때, target sentence로 번역될 $P(t|s)$를 modeling 하는 것

A basic form of NMT consist of two components:

- An encoder which computes a representations $\boldsymbol{s}$ for each source sentence
- A decoder which generates one target word at a time and hence decompose the conditional probability as

$$\log p(y|x) = \sum_{t=1}^{m} \log p(y_t|y_{<t}, \boldsymbol{s})$$

$$p(y_t|y_{<t}, \boldsymbol{s}) = softmax\big(g(\boldsymbol{h}_t)\big), \boldsymbol{h}_t = f\big(\boldsymbol{h}_{j-1}, \boldsymbol{s}\big)$$

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

$$source\ sentence : x_1, \dots, x_n$$

$$target\ sentence : y_1, \dots, y_m$$

Attention-based Models

일반적으로 neural net을 training하는 관점에서의 attention은 서로 다른 두 modality의 alignment를 neural net이 배우게 하는 것

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)

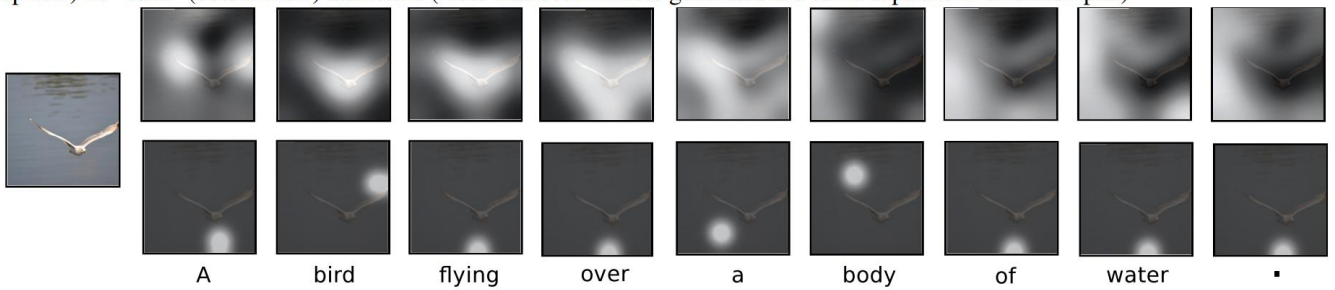
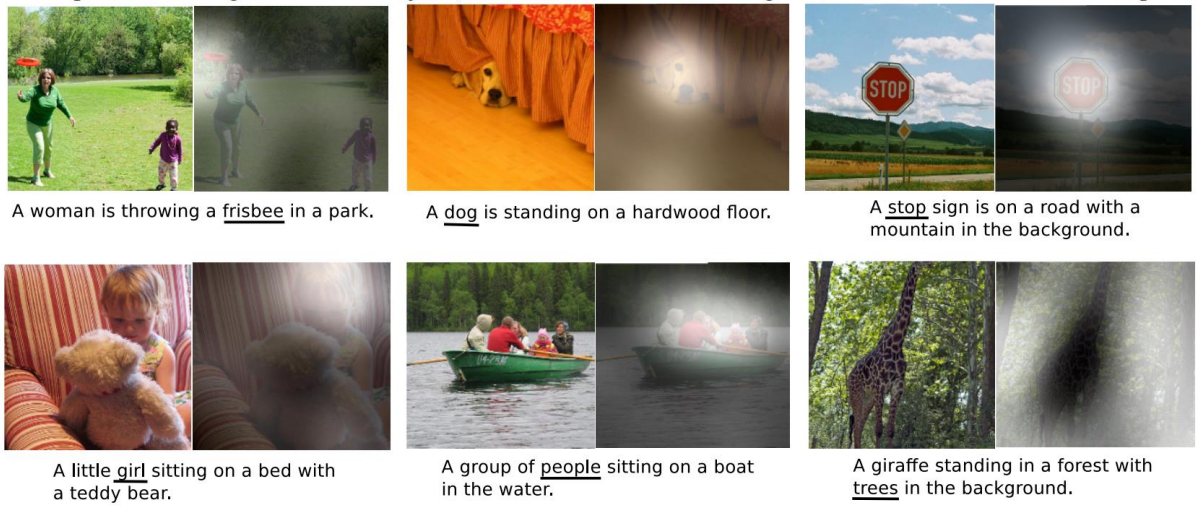


Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (https://arxiv.org/abs/1502.03044)

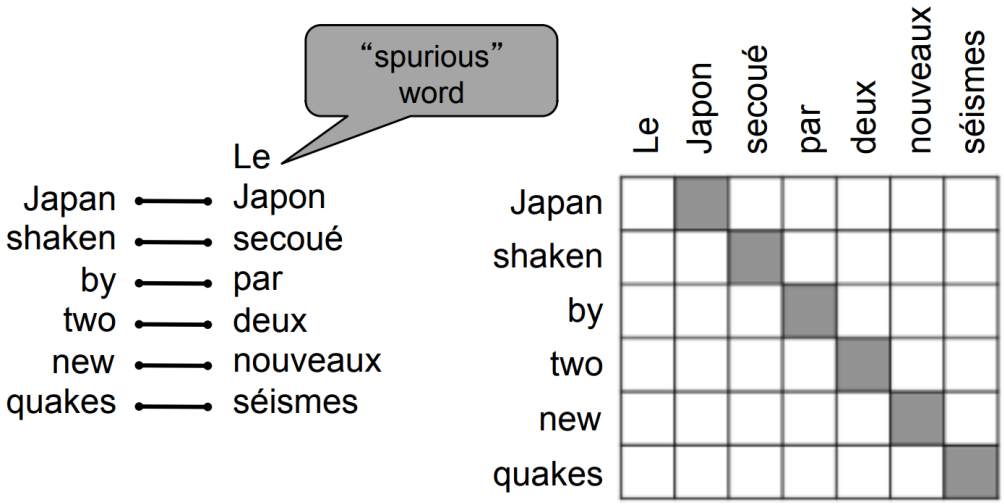
Attention-based Models

Neural Machine Translation에서의 attention은 source sentence와 target sentence의 alignment를 neural net이 배우게 하는 것

What is alignment?

Alignment is the correspondence between particular words in the translated sentence pair.

- Note: Some words have no counterpart



http://web.stanford.edu/class/cs224n/lectures/lecture10.pdf

Attention-based Models

backbone으로 encoder, decoder가 stacked lstm인 구조를 활용하며, 해당 구조에 Attention 방법론을 추가

$\boldsymbol{c}_t$를 만드는 방식이 결국 논문에서 말하는 Global Attention과 Local Attention!

$$\log p(y|x) = \sum_{t=1}^{m} \log p(y_t|y_{<t}, \boldsymbol{s})$$

$$p(y_t|y_{<t}, \boldsymbol{s}) = softmax(\boldsymbol{W}\boldsymbol{h}_t),$$
$$\boldsymbol{h}_t = f(\boldsymbol{h}_{t-1}, \boldsymbol{s})$$

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

$source\ sentence:$
$x_1, \dots, x_n$

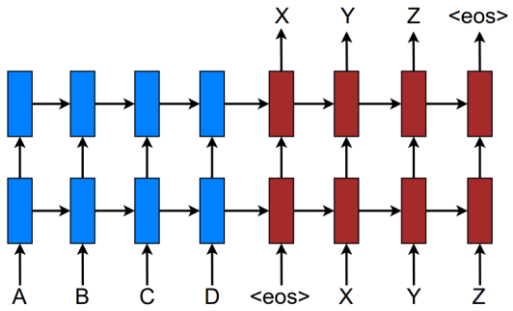
$target\ sentence:$
$y_1, \dots, y_m$

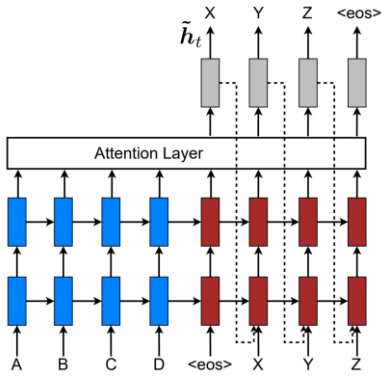
$$\log p(y|x) = \sum_{j=1}^{m} \log p(y_t|y_{<t}, \boldsymbol{c}_t)$$

$$p(y_t|y_{<t}, \boldsymbol{c}_t) = softmax(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{h}}_t),$$
$$\boldsymbol{h}_t = f(\boldsymbol{h}_{t-1}, \widetilde{\boldsymbol{h}}_{t-1}, \boldsymbol{s}), \widetilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_{\boldsymbol{c}}[\boldsymbol{c}_t; \boldsymbol{h}_t])$$

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

$source\ sentence: x_1, \dots, x_n$
$target\ sentence: y_1, \dots, y_m$





Attention-based Models

Global Attention

Global Attention의 idea는 encoder의 모든 step의 hidden state를 c_t를 만들 때 고려하자는 것

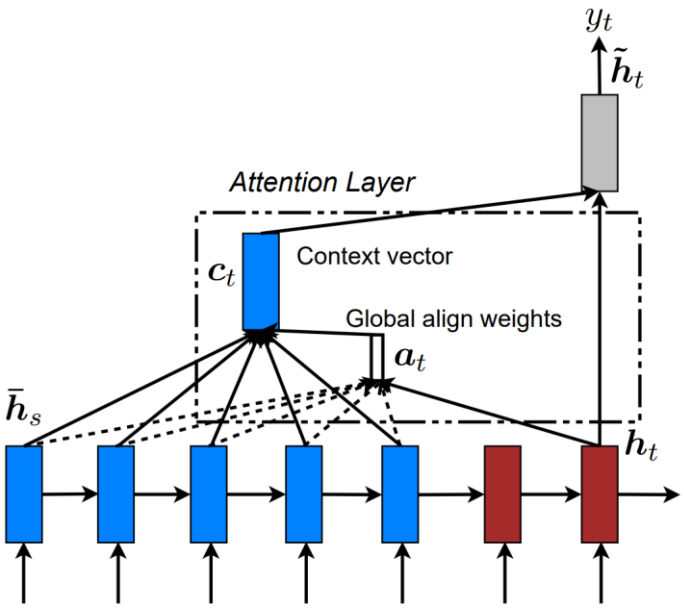


Figure 2: **Global attentional model** – at each time step t, the model infers a *variable-length* alignment weight vector $\boldsymbol{a}_t$ based on the current target state $\boldsymbol{h}_t$ and all source states $\bar{\boldsymbol{h}}_s$. A global context vector $\boldsymbol{c}_t$ is then computed as the weighted average, according to $\boldsymbol{a}_t$, over all the source states.

Location-based function

$$\boldsymbol{a_t} = \text{softmax}(\boldsymbol{W_a h_t}) \quad location$$

Content-based function

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

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^T \bar{\boldsymbol{h}}_s & dot \\ \boldsymbol{h}_t^T \boldsymbol{W_a} \bar{\boldsymbol{h}}_s & general \\ \boldsymbol{v}_a^T \tanh(\boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]) & concat \end{cases}$$

Attention-based Models

Local Attention (1/2)

Global attention의 경우 source sentence가 길어지면 translate하기가 impractical, expensive해지는 문제가 발생 → source sentence의 subset만 고려하는 것

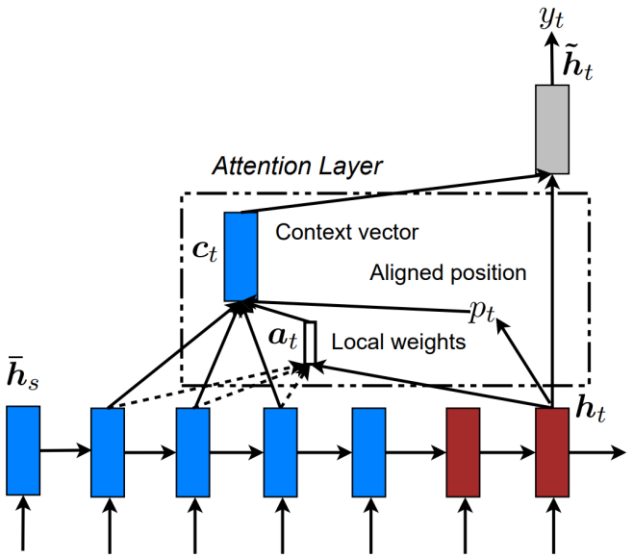


Figure 3: **Local attention model** – the model first predicts a single aligned position p_t for the current target word. A window centered around the source position p_t is then used to compute a context vector $\boldsymbol{c}_t$, a weighted average of the source hidden states in the window. The weights $\boldsymbol{a_t}$ are inferred from the current target state $\boldsymbol{h}_t$ and those source states $\bar{\boldsymbol{h}}_s$ in the window.

than the hard attention approach. In concrete details, the model first generates an aligned position p_t for each target word at time t. The context vector $\boldsymbol{c}_t$ is then derived as a weighted average over the set of source hidden states within the window $[p_t - D, p_t + D]$; D is empirically selected.[8] Unlike the global approach, the local alignment vector $\boldsymbol{a}_t$ is now fixed-dimensional, i.e., $\in \mathbb{R}^{2D+1}$. We consider two variants of the model as below.

Monotonic alignment (**local-m**) – we simply set $p_t = t$ assuming that source and target sequences are roughly monotonically aligned. The alignment vector $\boldsymbol{a}_t$ is defined according to Eq. (7).[9]

Attention-based Models

Local Attention (2/2)

Global attention의 경우 source sentence가 길어지면 translate하기가 impractical, expensive해지는 문제가 발생 → source sentence의 subset만 고려하는 것

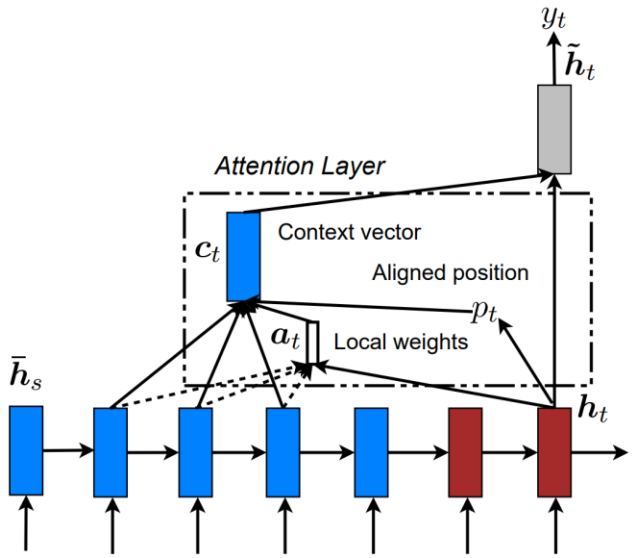


Figure 3: **Local attention model** – the model first predicts a single aligned position p_t for the current target word. A window centered around the source position p_t is then used to compute a context vector $\boldsymbol{c}_t$, a weighted average of the source hidden states in the window. The weights $\boldsymbol{a_t}$ are inferred from the current target state $\boldsymbol{h}_t$ and those source states $\bar{\boldsymbol{h}}_s$ in the window.

Predictive alignment (**local-p**) – instead of assuming monotonic alignments, our model predicts an aligned position as follows:

$$p_t = S \cdot \text{sigmoid}(\boldsymbol{v}_p^\top \tanh(\boldsymbol{W_p h}_t)), \quad (9)$$

$\boldsymbol{W_p}$ and $\boldsymbol{v}_p$ are the model parameters which will be learned to predict positions. S is the source sentence length. As a result of sigmoid, $p_t \in [0, S]$. To favor alignment points near p_t, we place a Gaussian distribution centered around p_t. Specifically, our alignment weights are now defined as:

$$\boldsymbol{a}_t(s) = \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right) \quad (10)$$

We use the same align function as in Eq. (7) and the standard deviation is empirically set as $\sigma = \frac{D}{2}$. Note that p_t is a *real* nummber; whereas s is an *integer* within the window centered at p_t.[10]

Attention-based Models

Input-feeding Approach

Global Attention, Local Attention은 time step마다 independent 하게 결정되므로 suboptimal → suboptimal을 최대한 줄이기 위해 이전 step의 attention 정보를 전달

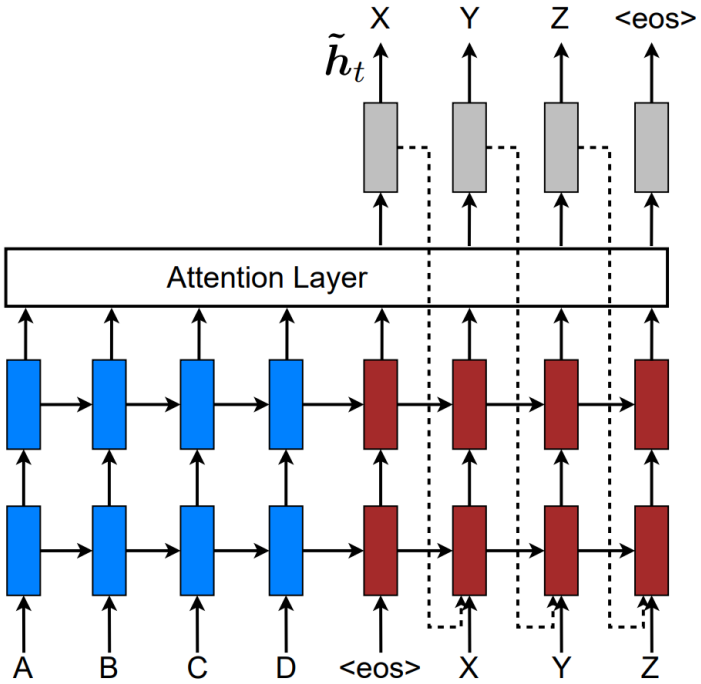


Figure 4: **Input-feeding approach** – Attentional vectors $\tilde{h}_t$ are fed as inputs to the next time steps to inform the model about past alignment decisions.

Experiments

English-German Results

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*)

Table 1: **WMT'14 English-German results** – shown are the perplexities (ppl) and the *tokenized* BLEU scores of various systems on newstest2014. We highlight the **best** system in bold and give *progressive* improvements in italic between consecutive systems. *local-p* referes to the local attention with predictive alignments. We indicate for each attention model the alignment score function used in pararentheses.

System	BLEU
SOTA – *NMT* + *5-gram rerank* (MILA)	24.9
Our ensemble 8 models + unk replace	**25.9**

Table 2: **WMT'15 English-German results** – *NIST* BLEU scores of the existing WMT'15 SOTA system and our best one on newstest2015.

Experiments

German-English Results

System	Ppl.	BLEU
WMT'15 systems		
SOTA – *phrase-based* (Edinburgh)		**29.2**
NMT + 5-gram rerank (MILA)		27.6
Our NMT systems		
Base (reverse)	14.3	16.9
+ global (*location*)	12.7	19.1 (+*2.2*)
+ global (*location*) + feed	10.9	20.1 (+*1.0*)
+ global (*dot*) + drop + feed	9.7	22.8 (+*2.7*)
+ global (*dot*) + drop + feed + unk		24.9 (+*2.1*)

Table 3: **WMT'15 German-English results** – performances of various systems (similar to Table 1). The *base* system already includes source reversing on which we add *global* attention, *drop*out, input *feed*ing, and *unk* replacement.

Analysis

Learning curves & Effects of Translating Long Sentences

제안한 Attention mechanism이 적용된 NMT가 그렇지 않은 것보다 학습의 수렴속도가 빠르며, long sentences를 처리하는데 효과적임을 알 수 있음

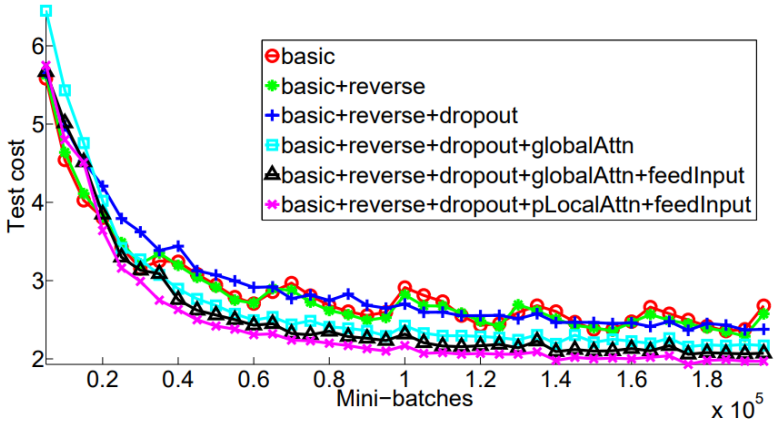


Figure 5: **Learning curves** – test cost (ln perplexity) on newstest2014 for English-German NMTs as training progresses.

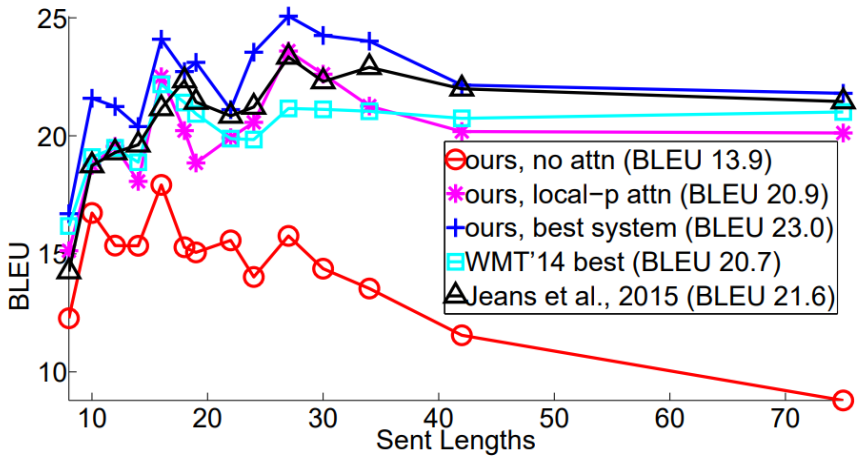


Figure 6: **Length Analysis** – translation qualities of different systems as sentences become longer.

Analysis

Choices of Attentional Architectures & Alignment Quality (1/2)

System	Ppl	BLEU	
		Before	After unk
global (location)	6.4	18.1	19.3 (+1.2)
global (dot)	6.1	18.6	20.5 (+1.9)
global (general)	6.1	17.3	19.1 (+1.8)
local-m (dot)	>7.0	x	x
local-m (general)	6.2	18.6	20.4 (+1.8)
local-p (dot)	6.6	18.0	19.6 (+1.9)
local-p (general)	**5.9**	**19**	**20.9 (+1.9)**

Table 4: **Attentional Architectures** – performances of different attentional models. We trained two local-m (dot) models; both have ppl $>$ 7.0.

Method	AER
global (location)	0.39
local-m (general)	0.34
local-p (general)	0.36
ensemble	0.34
Berkeley Aligner	0.32

Table 6: **AER scores** – results of various models on the RWTH English-German alignment data.

Analysis

Choices of Attentional Architectures & Alignment Quality (2/2)

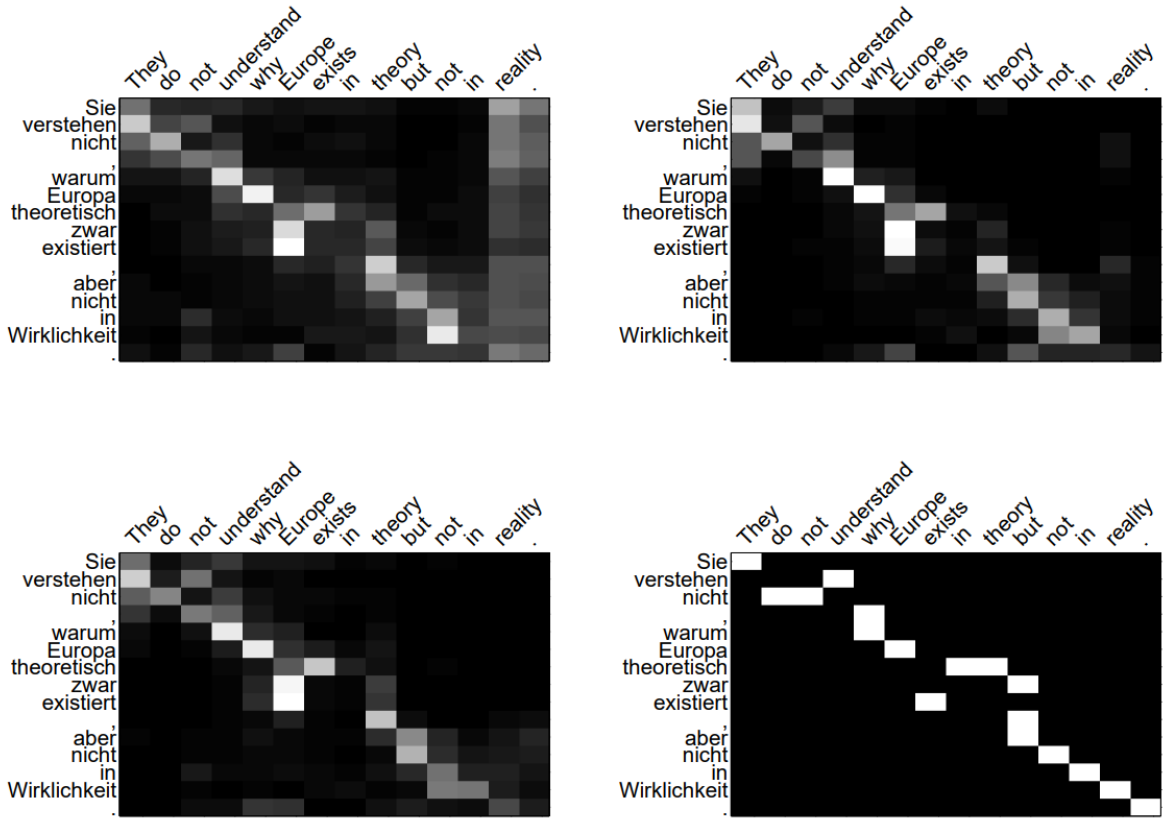


Figure 7: **Alignment visualizations** – shown are images of the attention weights learned by various models: (top left) global, (top right) local-m, and (bottom left) local-p. The *gold* alignments are displayed at the bottom right corner.

Analysis

Sample Translations

English-German translations	
src	Orlando Bloom and Miranda Kerr still love each other
ref	Orlando Bloom und *Miranda Kerr* lieben sich noch immer
best	Orlando Bloom und *Miranda Kerr* lieben einander noch immer .
base	Orlando Bloom und **Lucas Miranda** lieben einander noch immer .
src	″ We ′ re pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security , ″ said Roger Dow , CEO of the U.S. Travel Association .
ref	“ Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht ” , sagte *Roger Dow* , CEO der U.S. Travel Association .
best	″ Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit *unvereinbar* ist ″ , sagte *Roger Dow* , CEO der US - die .
base	″ Wir freuen uns über die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit ″ , sagte *Roger* **Cameron** , CEO der US - <unk> .
German-English translations	
src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben .
ref	However , in an interview , Bloom has said that he and *Kerr* still love each other .
best	In an interview , however , Bloom said that he and *Kerr* still love .
base	However , in an interview , Bloom said that he and **Tina** were still <unk> .
src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The *austerity imposed by Berlin and the European Central Bank , coupled with the straitjacket* imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
best	Because of the strict *austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket* in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	Because of the pressure **imposed by the European Central Bank and the Federal Central Bank with the strict austerity** imposed on the national economy in the face of the single currency , many people believe that the European project has gone too far .

Table 5: **Sample translations** – for each example, we show the source (*src*), the human translation (*ref*), the translation from our best model (*best*), and the translation of a non-attentional model (*base*). We italicize some *correct* translation segments and highlight a few **wrong** ones in bold.

Conclusion & QnA

NMT에 적용할 수 있는 attentional mechanism으로 Global approach와 Local approach를 제안하고, WMT'14, WMT'15에서 SOTA임을 확인