코리아IT아카데미

ADVISOR

Master. RYU

STUDENT

홍준표, 양문기,김동휘,
유아람,이정훈

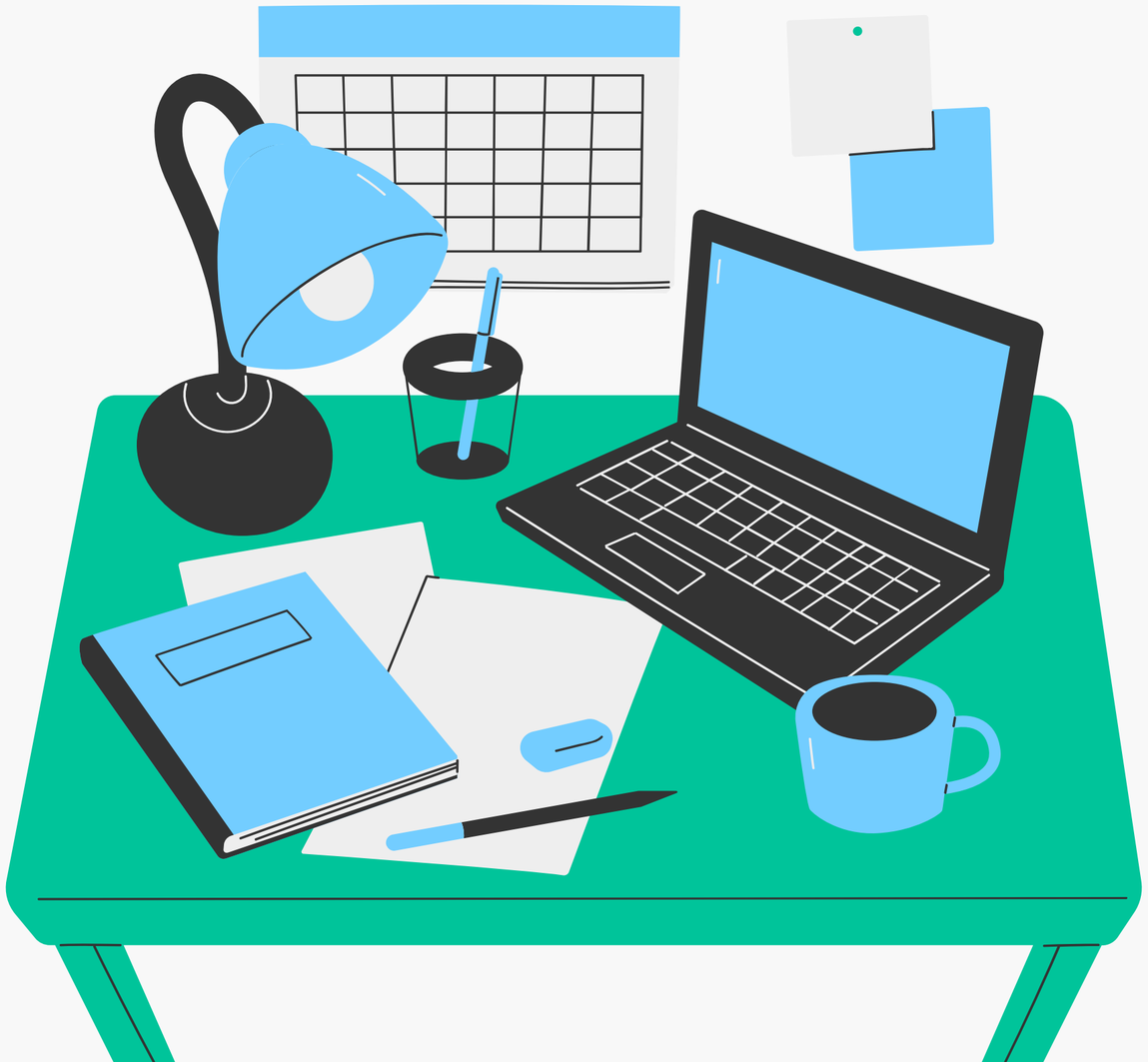
Attention Model 논문 분석

[Effective Approaches to Attention-based Neural Machine Translation]

목차

I Overview

- Better NMT
- 어텐션 메커니즘(Attention Mechanism)
- Global approach & Local approach
- English -> German



NMT 란

- Neural Machine Translation(신경망 기계 번역)
- End-to-End Learning
- 문맥 고려
- 장거리 종속성
- 다국어 번역 지원

Seq2Seq 란

- 입력 시퀀스를 다른 형태의 출력 시퀀스로 변환하는 딥러닝 모델 아키텍처
- 긴 문장 처리에 용이
- 가변 길이의 입출력을 위한 인코더/디코더 구조 채택
- 인코더가 출력하는 벡터 사이즈가 고정되어 있기 때문에 입력으로 들어오는 단어의 수가 매우 많아지면 성능이 떨어진다는 한계

▶ Seq2Seq 의 등장 배경

Attention Mechanism

- Soft Attention
- Hard Attention
- Global Attention(Soft)
- Local Attention(Soft + Hard)

III Global Attention Model - 1

- **Global Attention** : 입력 시퀀스의 모든 요소에 가중치를 계산 (Soft)
- **Local Attention** : 입력 시퀀스의 부분 요소에 가중치를 계산 (Soft + Hard)

모델이 입력 시퀀스의 각 요소에 동적으로 가중치를 부여하여 출력에 영향을 미친다.

현재 타겟 단어 y_t를 예측하는 데 도움이 되도록 소스 정보가 담긴 Context Vector(c_t)를 구하는 것이 목표

I can do this all day. -> I can do this ___ ___.

III Global Attention Model - 2

목표는 Context Vector를 구하는 것!!

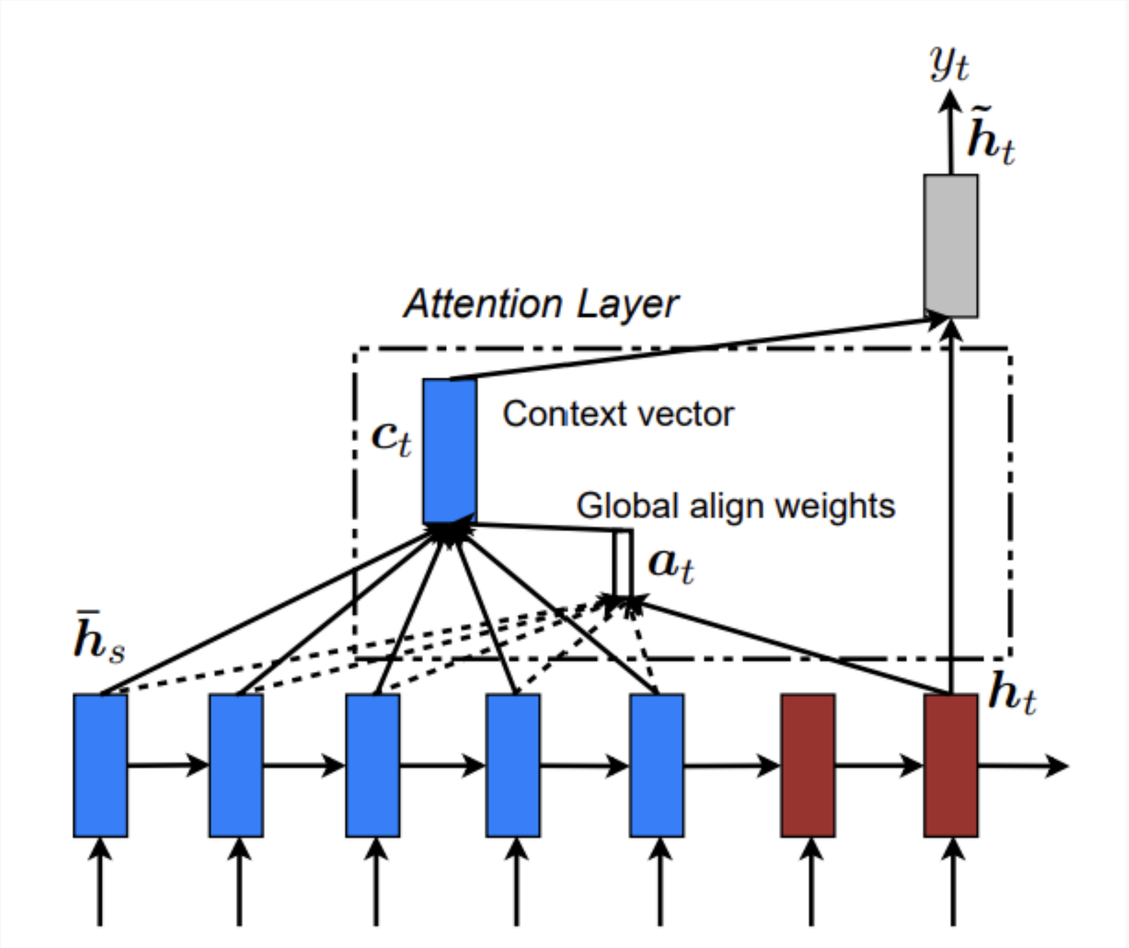
$$h_t \rightarrow a_t \rightarrow c_t \rightarrow \tilde{h}_t$$

- Input sequence의 모든 위치를 고려하여 output 예측.
- Alignment vector(at)는 중요도 정보를 담고있고 길이는 input sequence와 동일.

$$a_t(s) = align(h_t, \bar{h}_s)$$
$$= \frac{exp(score(h_t, \bar{h}_s))}{\sum_{\hat{s}} exp(score(h_t, \bar{h}_{\hat{s}}))}$$

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

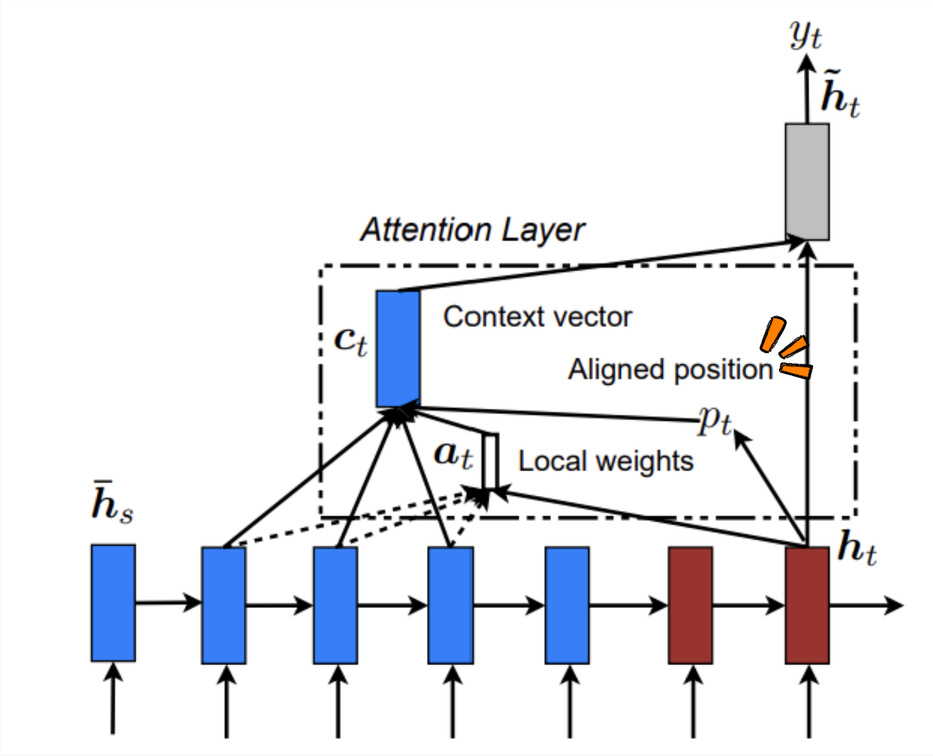
$$\boldsymbol{a}_t = \text{softmax}(\boldsymbol{W}_a \boldsymbol{h}_t) \qquad \textit{location}$$



IV Local Attention Model - 1

Global Attention은 입력 시퀀스의 길이가 길어질수록 모델의 계산 복잡도가 증가하고 메모리 사용량도 증가할 수 있다.

이러한 단점을 극복하기 위해 도입된것이 **Local Attention**



$$[p_t - D, p_t + D]$$

Context Vector의 차원 $2D + 1$

IV Local Attention Model - 2

Monotonic alignment (local-m)

$$p_t = t$$

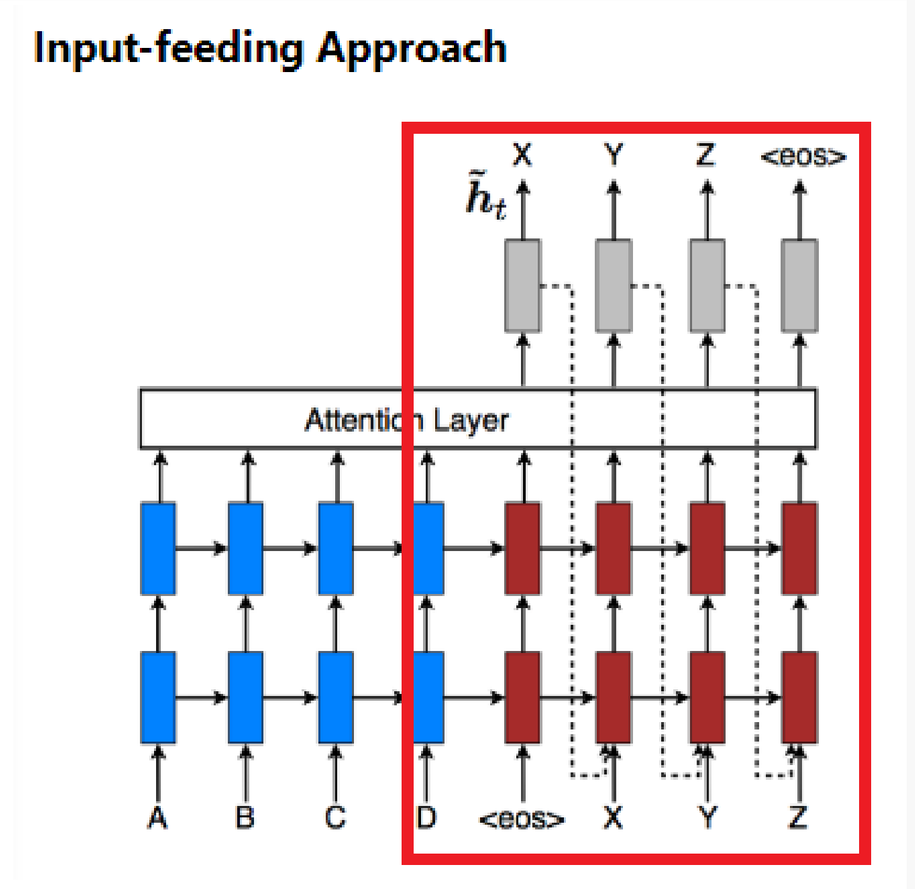
Predictive alignment (local-p)

$$p_t = S \cdot sigmoid(v_p^T tanh(W_p h_t))$$

I can do this all day. -> I can do this ___ ___.

Input Feeding Model

- 과거의 Alignment Information 이 지금의 Alignment Deisicion 에 고려되게 만들기 위한 방법
- Decoder 의 이전 출력을 현재 Step 의 입력으로 추가하거나 연결
- 병렬성이 침해되고 속도가 느려진다는 단점



VI Experiment & Result

□하이퍼파라미터 설정에 newstest2013 사용
테스트에 newstest2014, 2015 사용

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*)

사용된 평가 지표

tokenized12 BELU

NIST13 BELU

Base + reverse + dropout 보다 5.0 높음

VI Experiment & Result

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*)

✓ SOTA

기존 SOTA보다 1.4 BLEU 높음

VI Experiment & Result

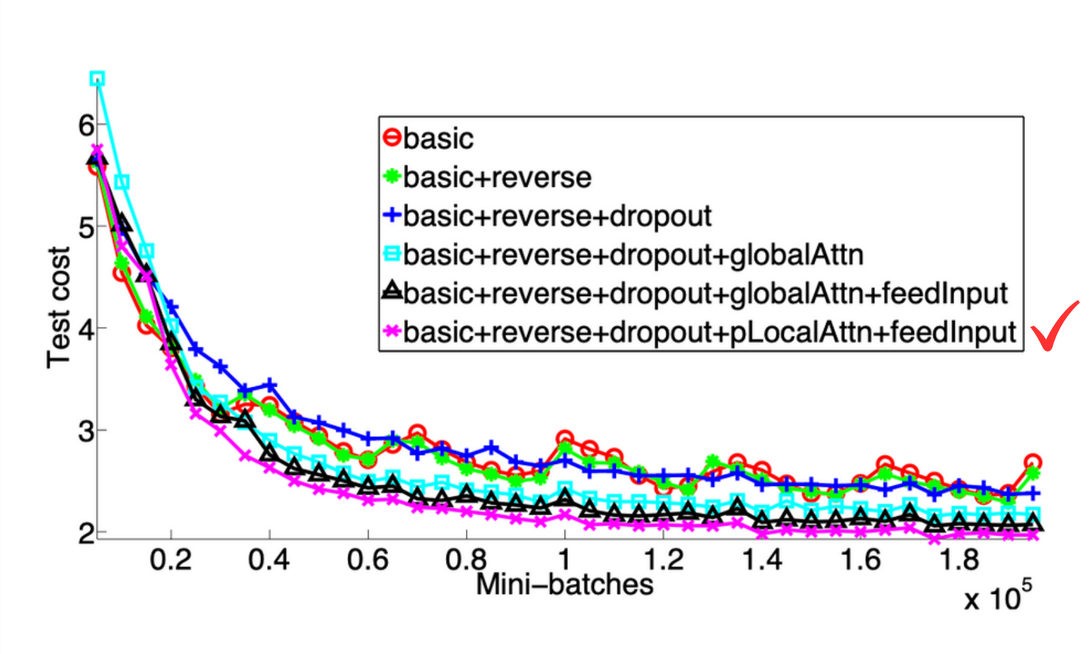
System	BLEU
Top – *NMT + 5-gram rerank* (Montreal)	24.9
Our ensemble 8 models + unk replace	**25.9**

WMT'15 English-German results

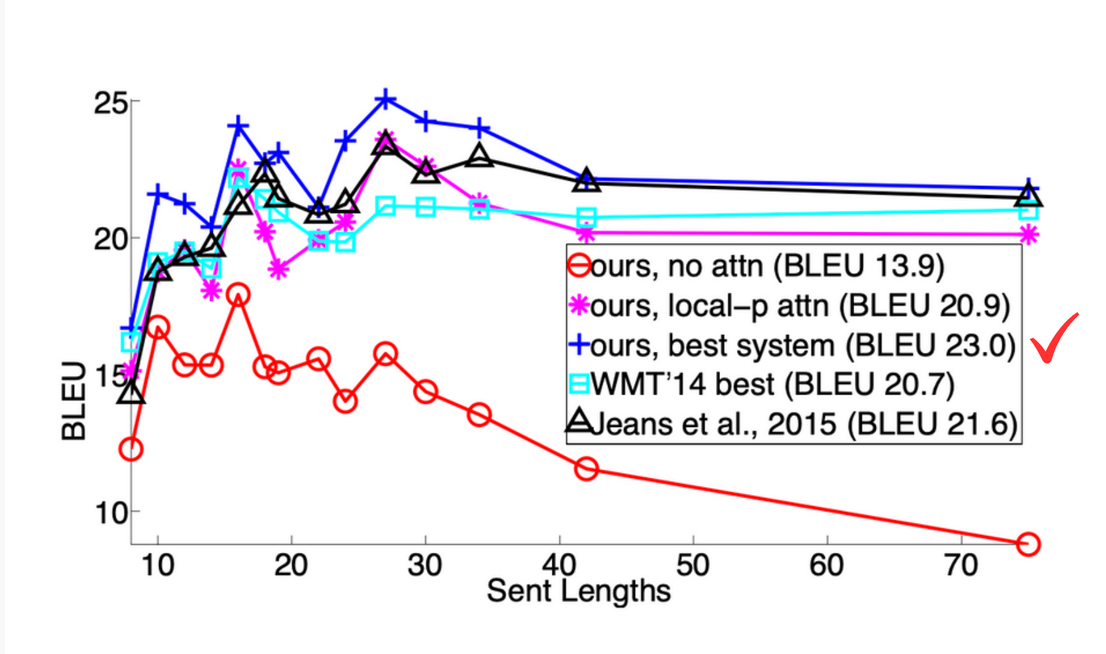
System	Ppl.	BLEU
WMT'15 systems		
SOTA – *phrase-based* (Edinburgh)		**29.2**
NMT + 5-gram rerank (MILA)		27.6
Our NMT systems		
Base (reverse)	14.3	16.9
+ global (*location*)	12.7	19.1 (+*2.2*)
+ global (*location*) + feed	10.9	20.1 (+*1.0*)
+ global (*dot*) + drop + feed	9.7	22.8 (+*2.7*)
+ global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

WMT'15 German-English results

WMT'14 English-German의 결과에 대해서만 분석



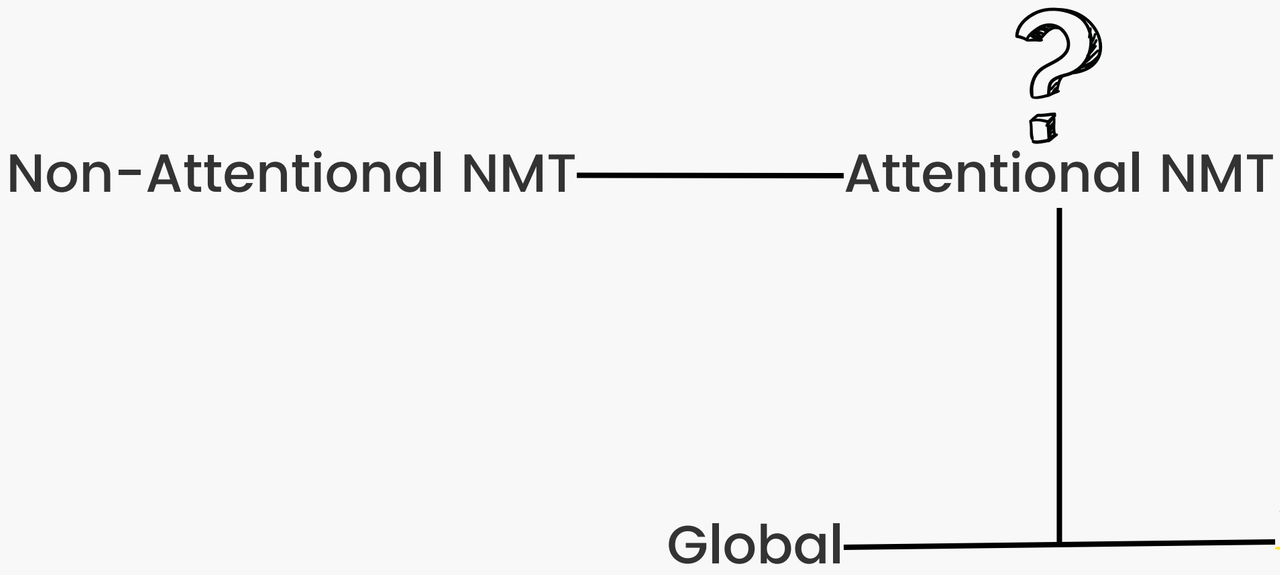
test 데이터에 대한 성능



문장 길이에 따른 BLEU

VII Conclusion & Discussions

English -> German



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)**

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

+dropout

+input-feeding

Thank you for listening!

참고