Syntax-Directed Attention for Neural Machine Translation

Kehai Chen[1], **Rui Wang**[2], Masao Utiyama[2], Eiichiro Sumita[2] and Tiejun Zhao[1]
[1]Harbin Institute of Technology, Harbin, China
[2]National Institute of Information and Communications Technology, Kyoto, Japan
{khchen, tjzhao}@hit.edu.cn, {wangrui, mutiyama, eiichiro.sumita}@nict.go.jp

1. Introduction

- In attention mechanism (Fig 1.a), alignment weights of the current target word often decrease to the left and right by linear distance (Fig 1.b) centering on the aligned source position and neglect syntax distance constraints.
- In linear distance, syntax-related source words are often far away from the aligned source word, and thus they can not be adequately taken into account during learning context vector.

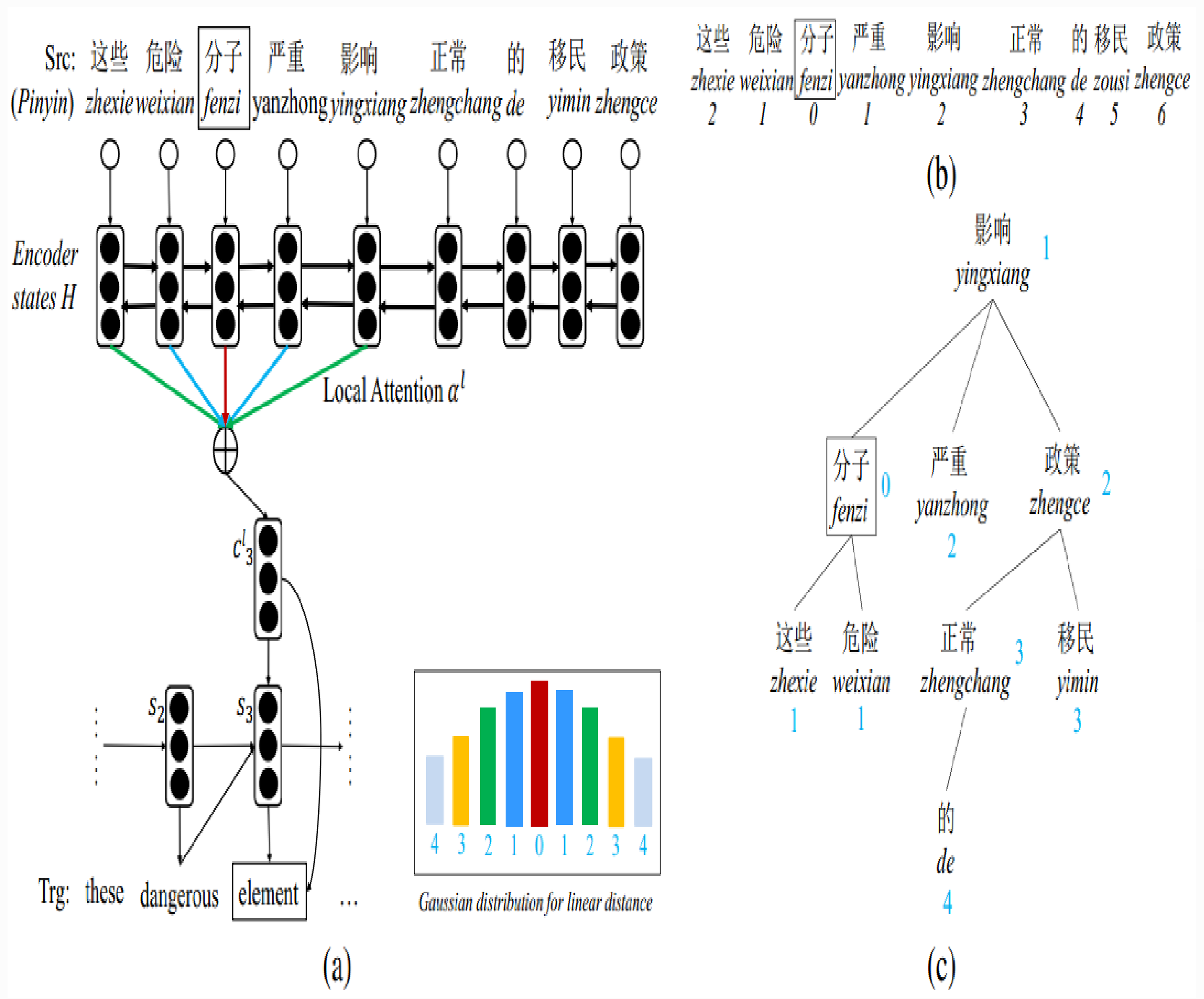

Fig 1: (a) NMT with the local attention. (b) Linear distance of word "fenzi". (c) Syntax distance of word "fenzi".from source dependency tree

- **Syntax-directed attention is proposed to capture syntax related source words with the predicted target word by syntax distance constraint (Fig 1.c) instead of linear distance constraints.**

2. Syntax Distance Constraint

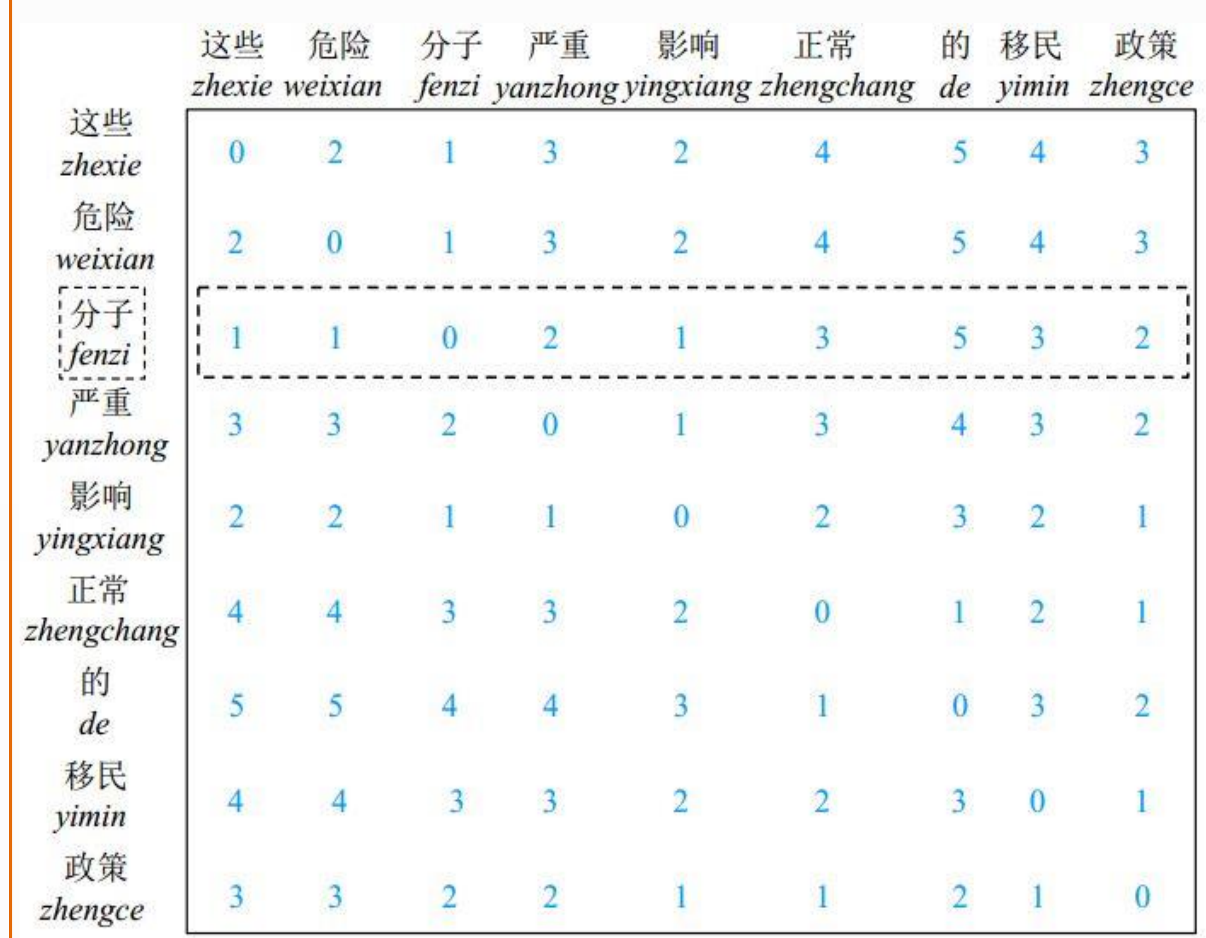

Fig 2: Syntax distance constraint mask matrix M for the dependency-based Chinese sentence in Fig 1.c, in which each row denotes the syntax distance mask of one source word, for example the dotted black box is syntax distance constraint mask for source word "*fenzi*".

Linear distance *of "fenzi" : {2, 1, 0, 1, 2, 3, 4, 5, 6},*
Syntax distance *of "fenzi" : {1, 1, 0, 2, 1. 3, 5, 3, 2}*

Given a source sentence X with dependency tree T, each node denotes a source word x_j and the distance between two connected nodes is defined as *one*. We then traverse each word in turn, and compute distances of all remaining words to the current traversed word x_j as its syntax distance constraint mask m_j. Finally, we learn a sequence of syntax distance constraint mask $\{m_0, m_1, \ldots, m_J\}$, which is a $J * J$ matrix *M:*

$$M = \{[m_1], [m_2], \ldots,, [m_J]\}.$$

3. Syntax-Directed Attention & Double Context

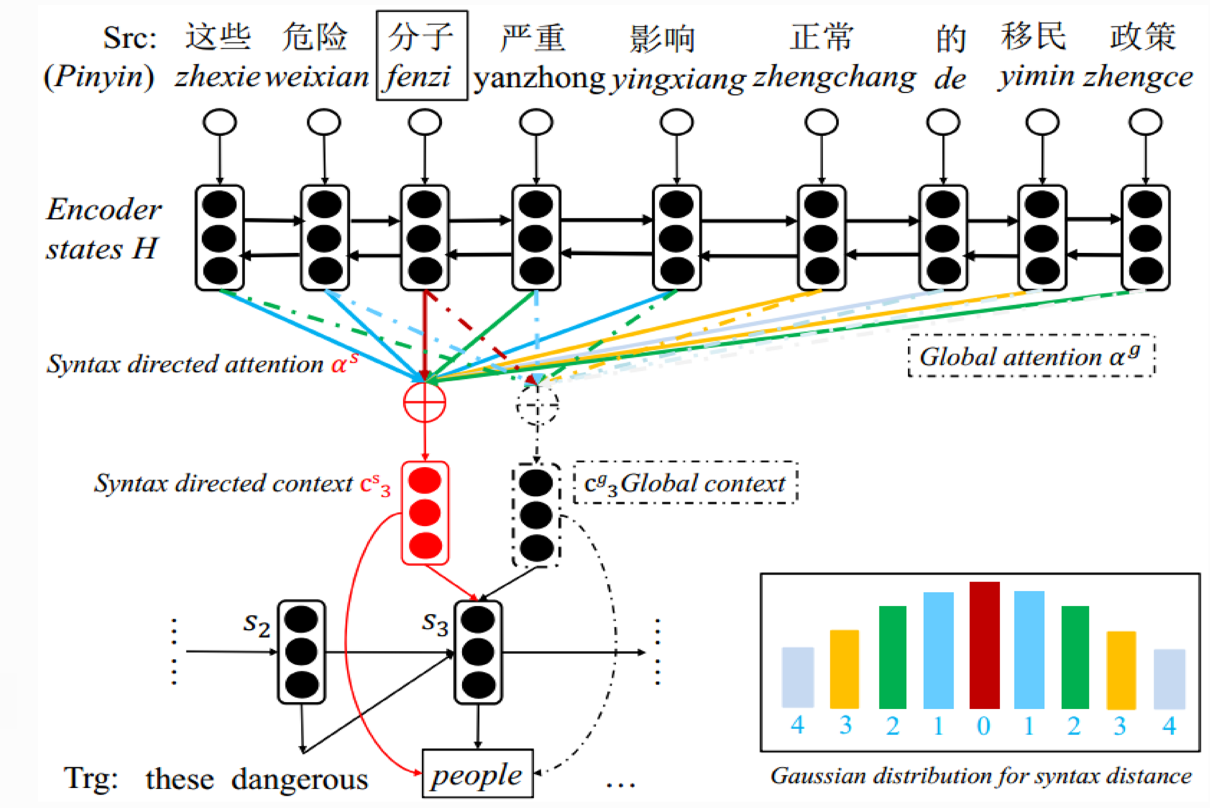

Fig 3: Syntax-directed attention and Double Context Mechanism (+dotted line)

Alignment score e^s_{ij} with $M[\mathrm{p}_i]$ is:

$$e^s_{ij} = \boldsymbol{v}\tanh\left(\boldsymbol{U_a s'_i} + \boldsymbol{W_a h_j}\right)\exp\left(-\frac{(M[p_i][j])^2}{2\sigma^2}\right)$$

$$p_i = J \cdot \mathrm{sigmoid}\left(\boldsymbol{v}'\tanh\left(\boldsymbol{W_p s'_i}\right)\right),$$

where p_i is source aligned position and $\boldsymbol{s'_i}$ is hidden state proposal.

The syntax-directed attention $\alpha^{s_n}_{ij}$ is normalized within n-gram distance:

$$\alpha^{s_n}_{ij} = \begin{cases} \frac{\exp(e^s_{ij})}{\sum_{k\in M[p_i][j]\le n}\exp(e^s_{ik})}, & j \in [p_i - n, p_i - n] \\ 0, & j \in [p_i - n, p_i - n] \end{cases}.$$

The syntax context vector $\boldsymbol{c}^s_i$:

$$\boldsymbol{c}^s_i = \sum_{j=0}^{J} \alpha^{s_n}_{ij}\boldsymbol{h_j}.$$

and thus the probability of the next word y_i is:

$$P(y_i|y_{<i}, X, T) = softmax\left(\boldsymbol{L}_o tanh(\boldsymbol{L_w E}[y_{i-1}] + \boldsymbol{L}_{cs}\boldsymbol{c}^s_i + \boldsymbol{L}_d\boldsymbol{s}_i)\right).$$

To further improve translation, we integrate a linear context vector $\boldsymbol{c}^g_i$ into syntax-directed attention to predict target word:

$$P(y_i|y_{<i}, X, T) = softmax\left(\boldsymbol{L}_o tanh\left(\boldsymbol{L}_w\boldsymbol{E}[y_{i-1}] + \boldsymbol{L}_{cg}\boldsymbol{c}^g_i + \boldsymbol{L}_{cs}c^s_i + \boldsymbol{L}_d\boldsymbol{s}_i\right)\right)$$

4.Experiments

ZH-EN	Dev (NIST02)	NIST03	NIST04	NIST05	NIST06	NIST08	AVG
PBSMT	33.15	31.02	33.78	30.33	29.62	23.53	29.66
GlobalAtt	37.12	35.24	37.49	34.60	32.48	26.32	33.23
Chen et al. (2017)	37.42	35.98	38.34	35.28	33.58	27.23	34.08
LocalAtt	37.31	35.57	37.85	34.93	32.74	26.83	33.58
FlexAtt	37.19	35.46	37.81	34.76	32.83	26.71	33.51
SDAtt	**38.01**	**36.67****†	**38.66****†	**35.74****†	**34.03****†	**27.66****†	**34.55**

EN-DE	Dev (newstest2012)	newstest2013	newstest2014	newstest2015	AVG
PBSMT	14.89	16.75	15.19	16.84	16.35
GlobalAtt	17.09	20.24	18.67	19.78	19.56
Chen et al. (2017)	17.48	21.03	19.43	20.56	20.31
LocalAtt	17.19	20.74	19.00	20.15	19.96
FlexibleAtt	17.24	20.57	19.12	20.03	19.91
SDAtt	**17.86**	**21.71****†	**20.36****†	**21.57****†	**21.21**

Table 1: Results on ZH-EN and EN-DE translation tasks for the proposed **syntax-directed attention**

ZH-EN	Dev (NIST02)	NIST03	NIST04	NIST05	NIST06	NIST08	AVG
PBSMT	33.15	31.02	33.78	30.33	29.62	23.53	29.66
GlobalAtt	37.12	35.24	37.49	34.60	32.48	26.32	33.23
+Chen et al. (2017)	38.11	37.35	39.00	36.12	33.78	27.81	34.81
+LocalAtt	37.89	37.06	38.73	36.10	33.62	27.43	34.59
+FlexibleAtt	37.97	36.86	38.56	35.62	33.94	27.37	34.47
+SDAtt	**38.61**	**38.19****†	**39.81****†	**36.74****	**34.63****†	**28.61****†	**35.60**

EN-DE	Dev (newstest2012)	newstest2013	newstest2014	newstest2015	AVG
PBSMT	14.89	16.75	15.19	16.84	16.35
GlobalAtt	17.09	20.24	18.67	19.78	19.56
+Chen et al. (2017)	18.03	21.44	19.96	21.07	20.82
+LocalAtt	17.78	21.26	19.87	20.67	20.6
+FlexibleAtt	17.56	21.10	19.76	20.74	20.53
+SDAtt	**18.65**	**22.11****†	**20.75****†	**22.05****†	**21.64**

Table 2: Results on ZH-EN and EN-DE translation tasks for the proposed **double context mechanism**

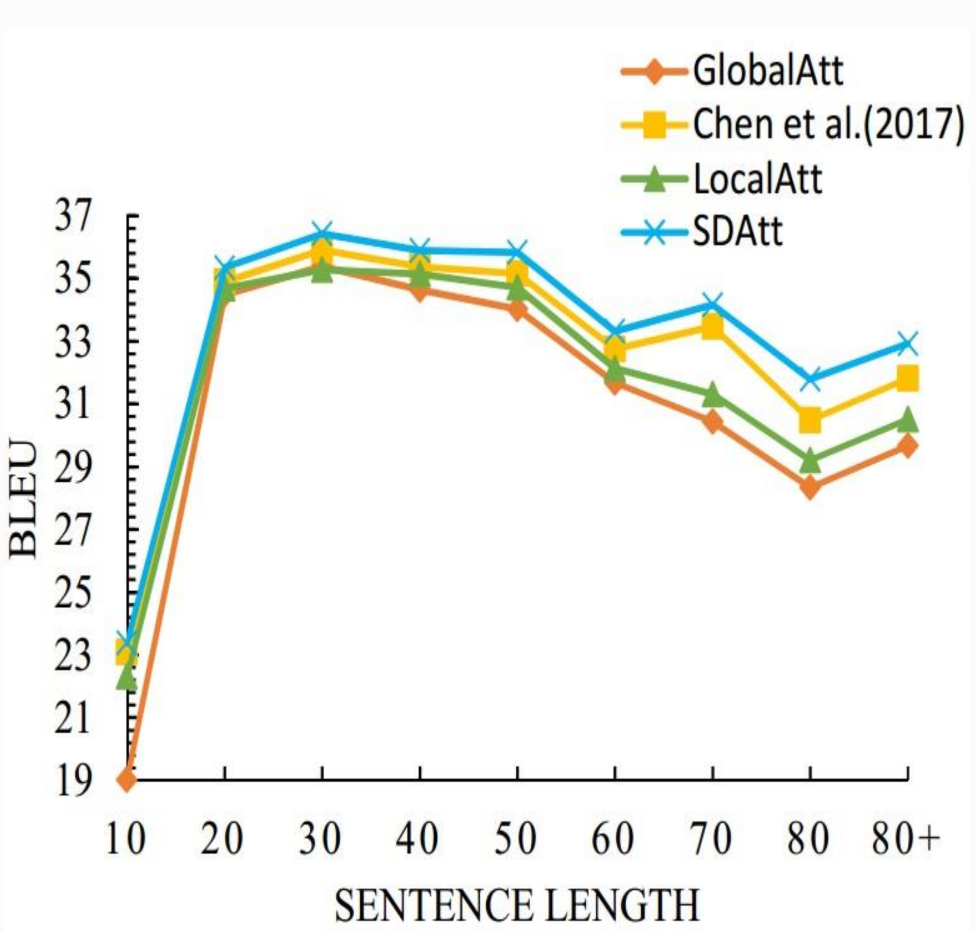

Fig 4: Translation qualities of different sentence lengths for SDAtt on the ZH-EN task

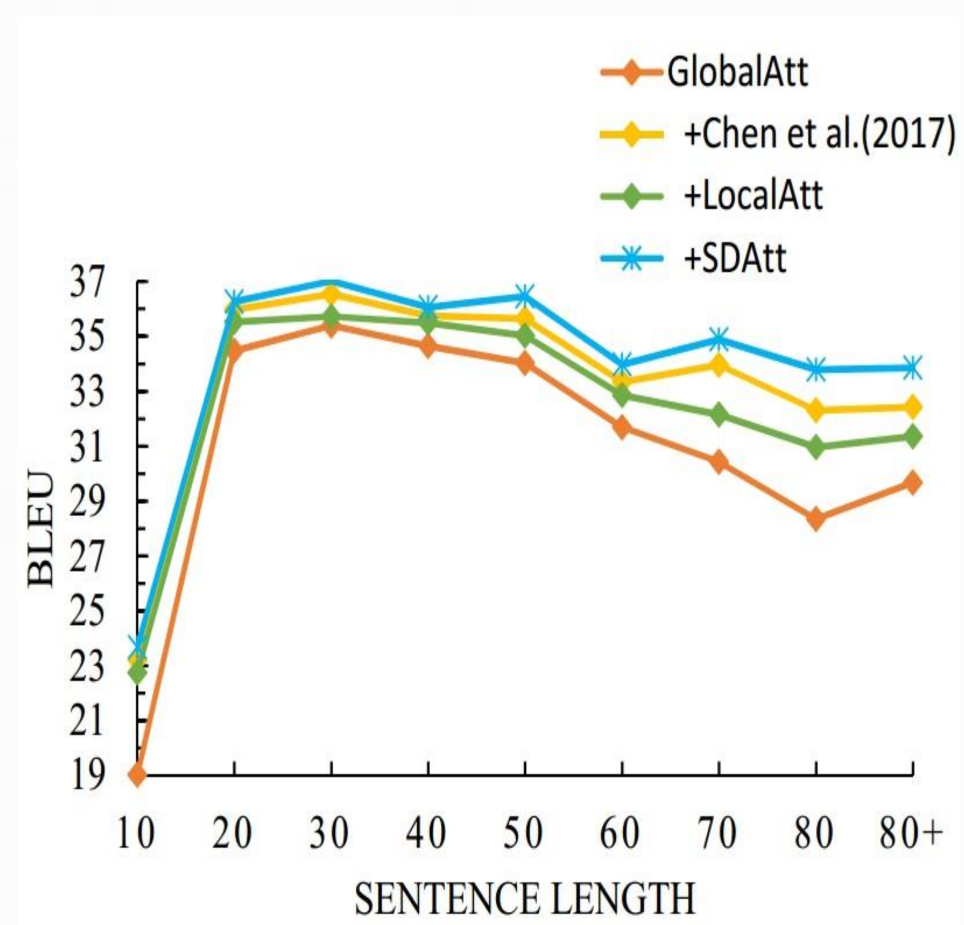

Fig 5: Translation qualities of different sentence lengths for GlobalAtt+SDAtt on the ZH-EN task

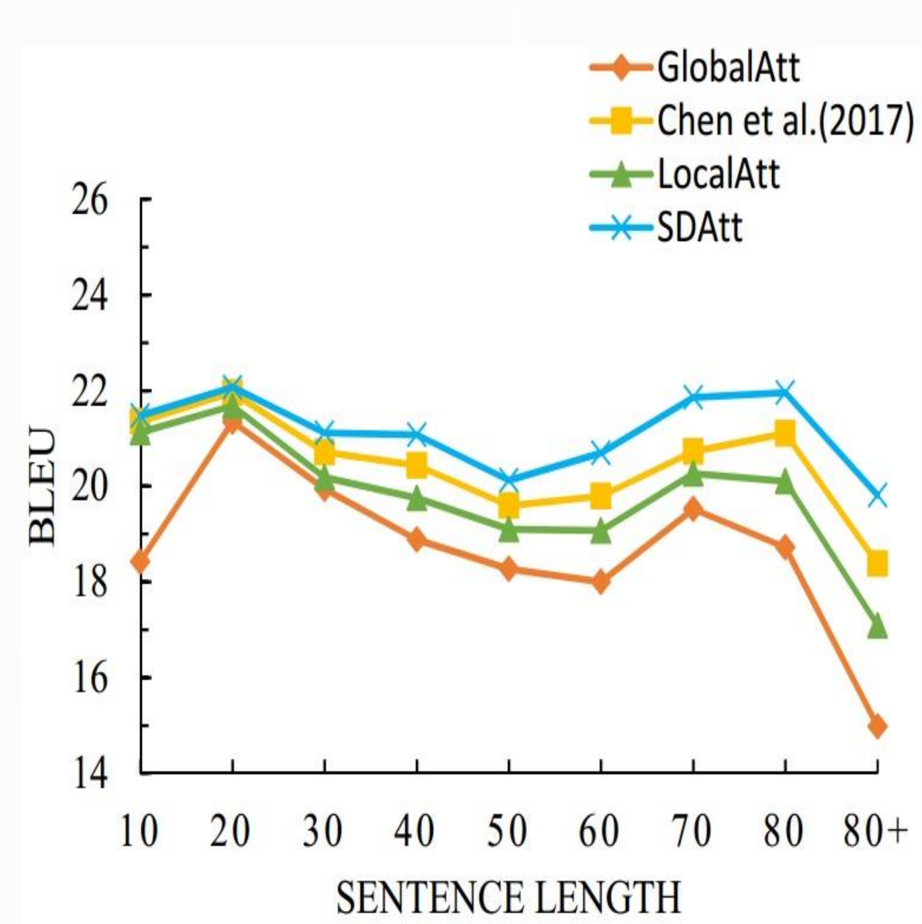

Fig 6: Translation qualities of different sentence lengths for SDAtt on the EN-DE task

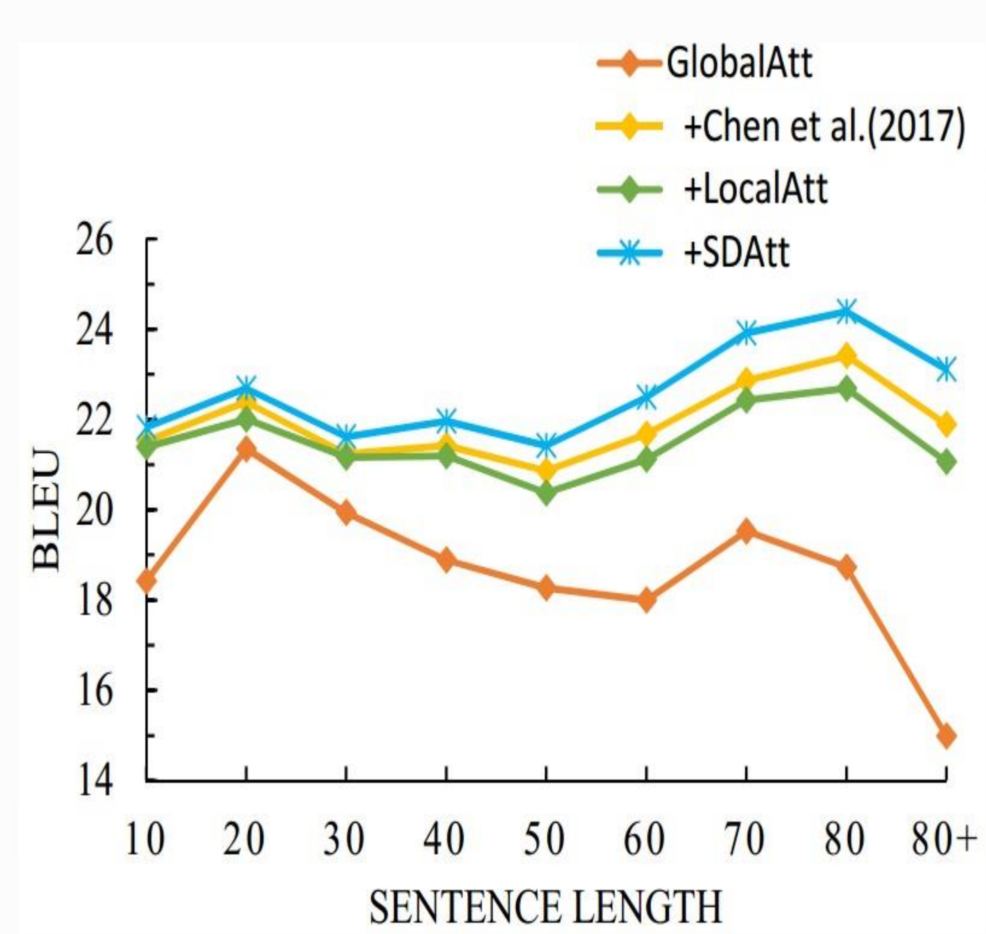

Fig 7: Translation qualities of different sentence lengths for GlobalAtt+SDAtt on the EN-DE task