Neural Machine Translation with Source Dependency Representation

Kehai Chen[1], Rui Wang[2], Masao Utiyama[2], Lemao Liu[3], Akihiro Tamura[4], Eiichiro Sumita[2] and Tiejun Zhao[1]

[1]*Harbin Institute of Technology, China*

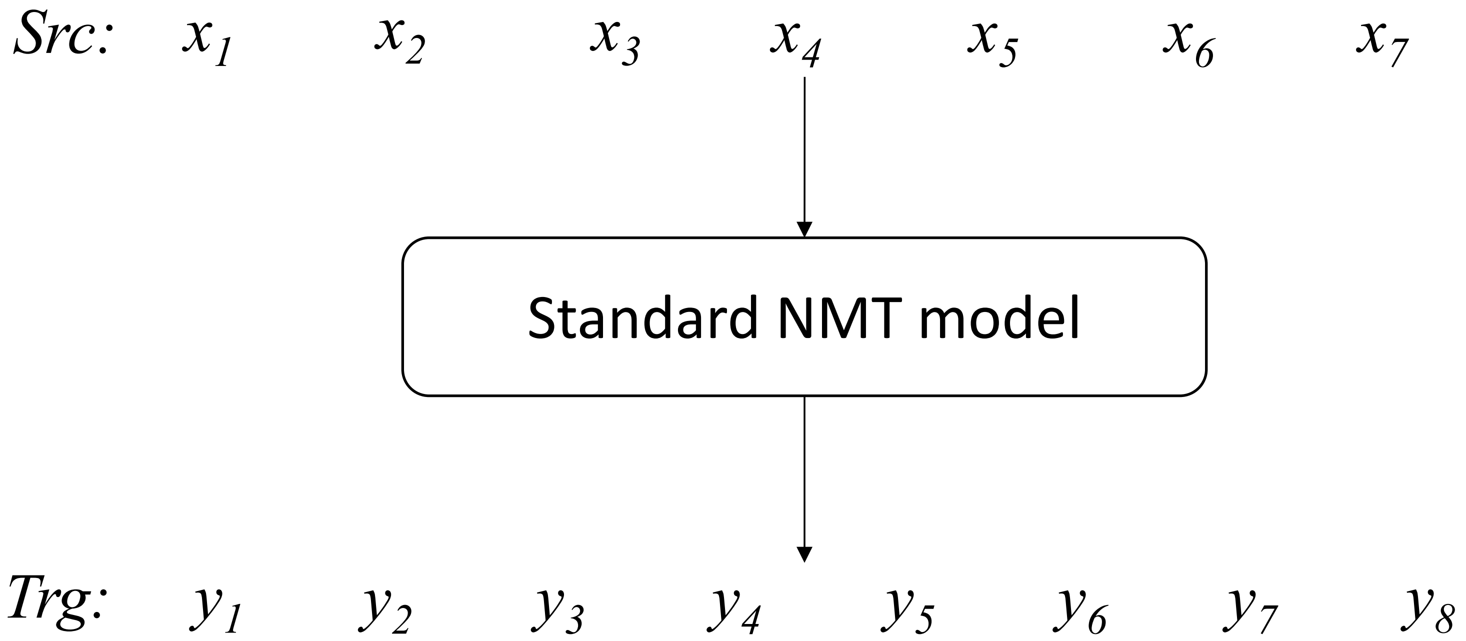
[2]*National Institute of Information and Communications Technology, Japan*

[3]*Tencent AI Lab, China*

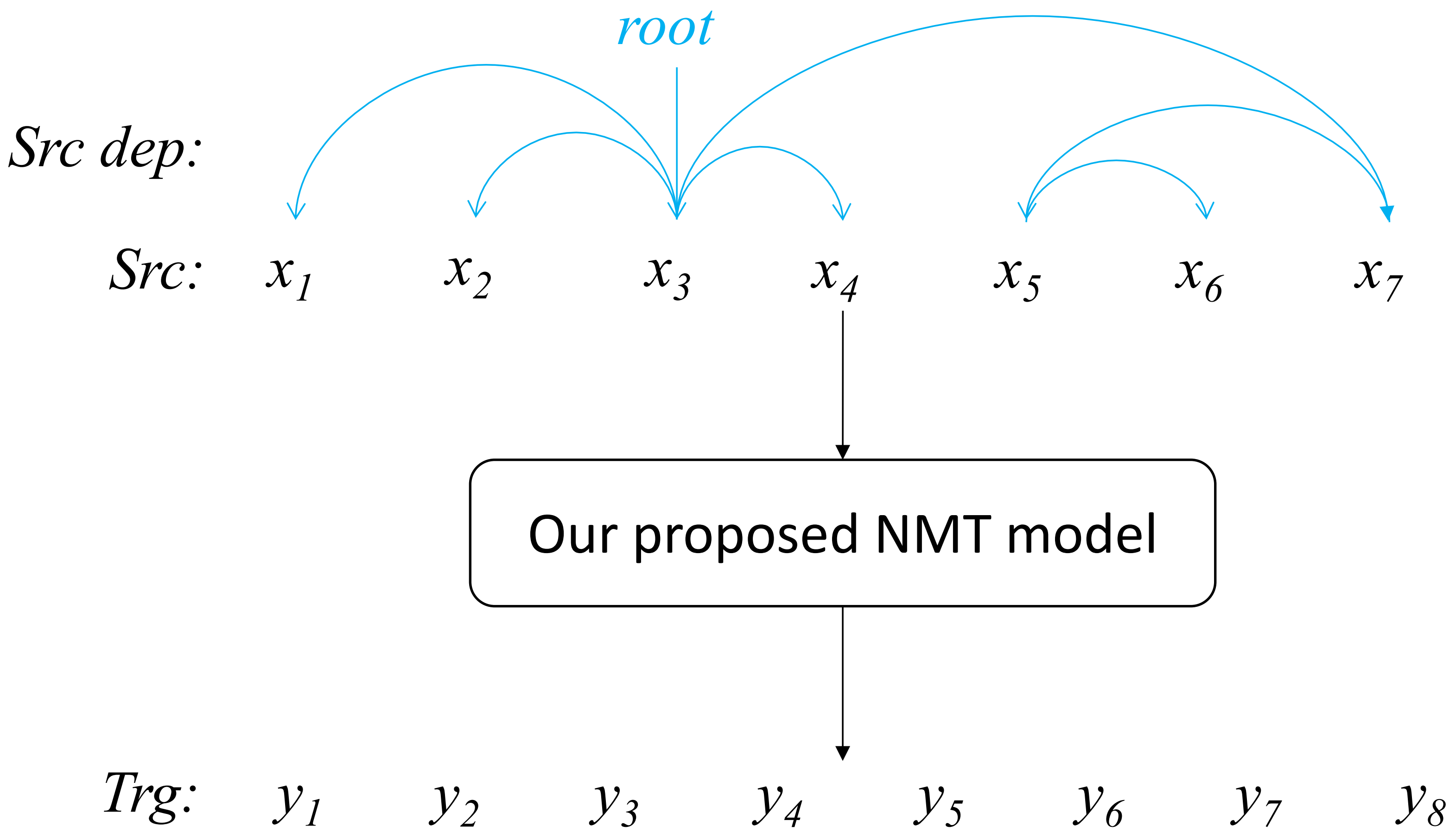
[4]*Ehime University, Japan*

Overview

- Traditional NMT Model



Overview



- Our proposed NMT model

Inspired by the syntax knowledge in SMT, we want to explicitly integrate source dependency information into NMT

Related Work

- NMT with source syntax information

  -Tree2seq (Eriguchi et al., 2016; Li et al., 2017; +other)
    Tree-based neural network is used to encode source phrase structures

  -Extending source inputs with syntax labels (Sennrich et al., 2016; Chen et al., 2017; +other)
    Dependency labels are concatenated to source word

Related Work

- NMT with source syntax information

  -Tree2seq (Eriguchi et al., 2016; Li et al., 2017; +other)
  Tree-based neural network is used to encode source phrase structures

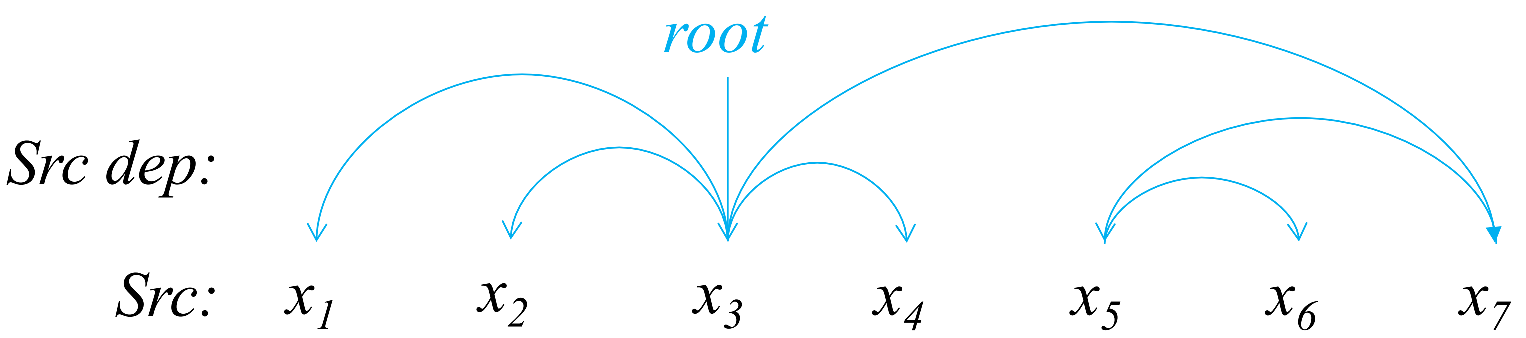
  -Extending source inputs with syntax labels (Sennrich et al., 2016; Chen et al., 2017; +other)
  Dependency labels are concatenated to source word

- Our work

  -A compromise between the two kinds of works

  -A novel double context approach to utilizing source dependency constraints

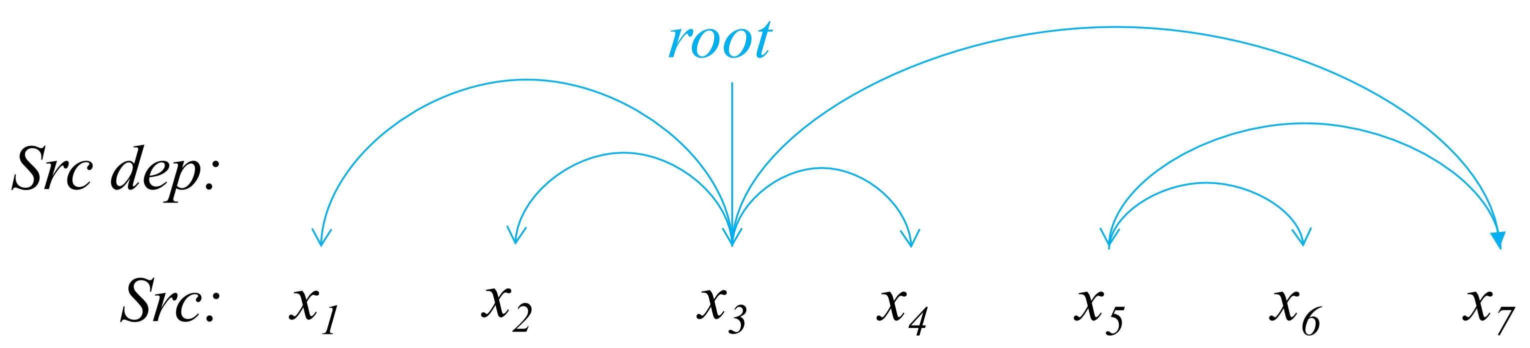
Source Dependency Representation (SDR)



- Extracting a dependency unit for each source word to capture source long-distance dependency constraints:

$$\boxed{U_j = < PA_{x_j}, SI_{x_j}, CH_{x_j} >}$$

Source Dependency Representation (SDR)



- Extracting a dependency unit for each source word to capture source long-distance dependency constraints:

$$\boxed{U_j = < PA_{x_j}, SI_{x_j}, CH_{x_j} >}$$

Where PA_{xj}, SI_{xj}, and CH_{xj} denote the parent, siblings and children words of source word x_j in a dependency structure.

Take x_2 as an example: $PA_{x_2} = < x_3 >$, *then,* $\boxed{U_2 = < x3, x1, x4, x7, \varepsilon >}$

$SI_{x_2} = < x_1, x_4, x_7 >$,

$CH_{x_2} = < \varepsilon >$,

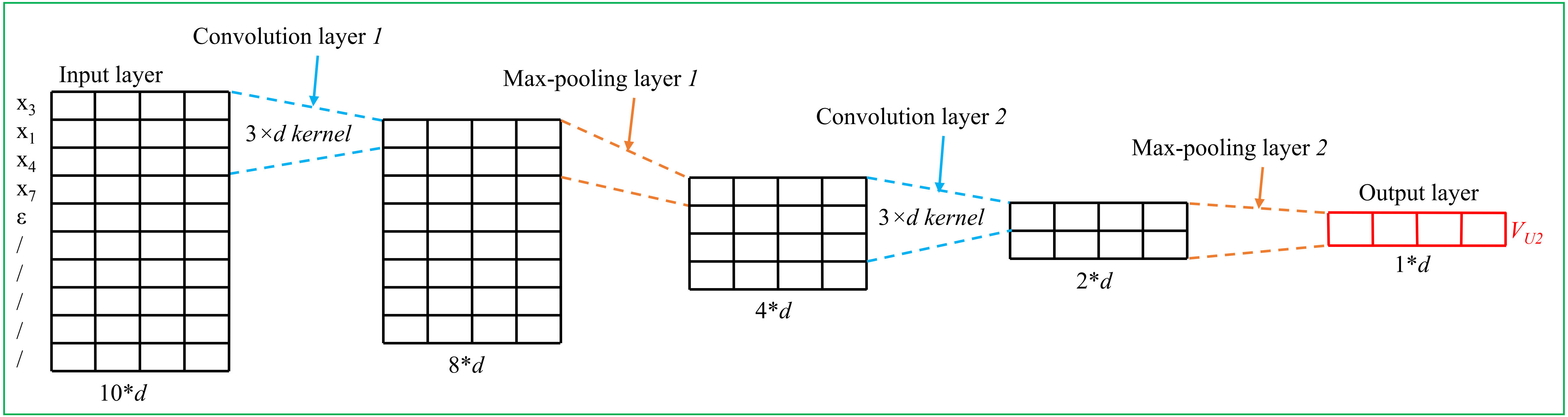
Source Dependency Representation (SDR)

- Learn sematic representation of each dependency unit

Take x_2 as an example: $PA_{x_2} =< x_3 >,$ then, $\boxed{U_2 =< x3, x1, x4, x7, \varepsilon >}$

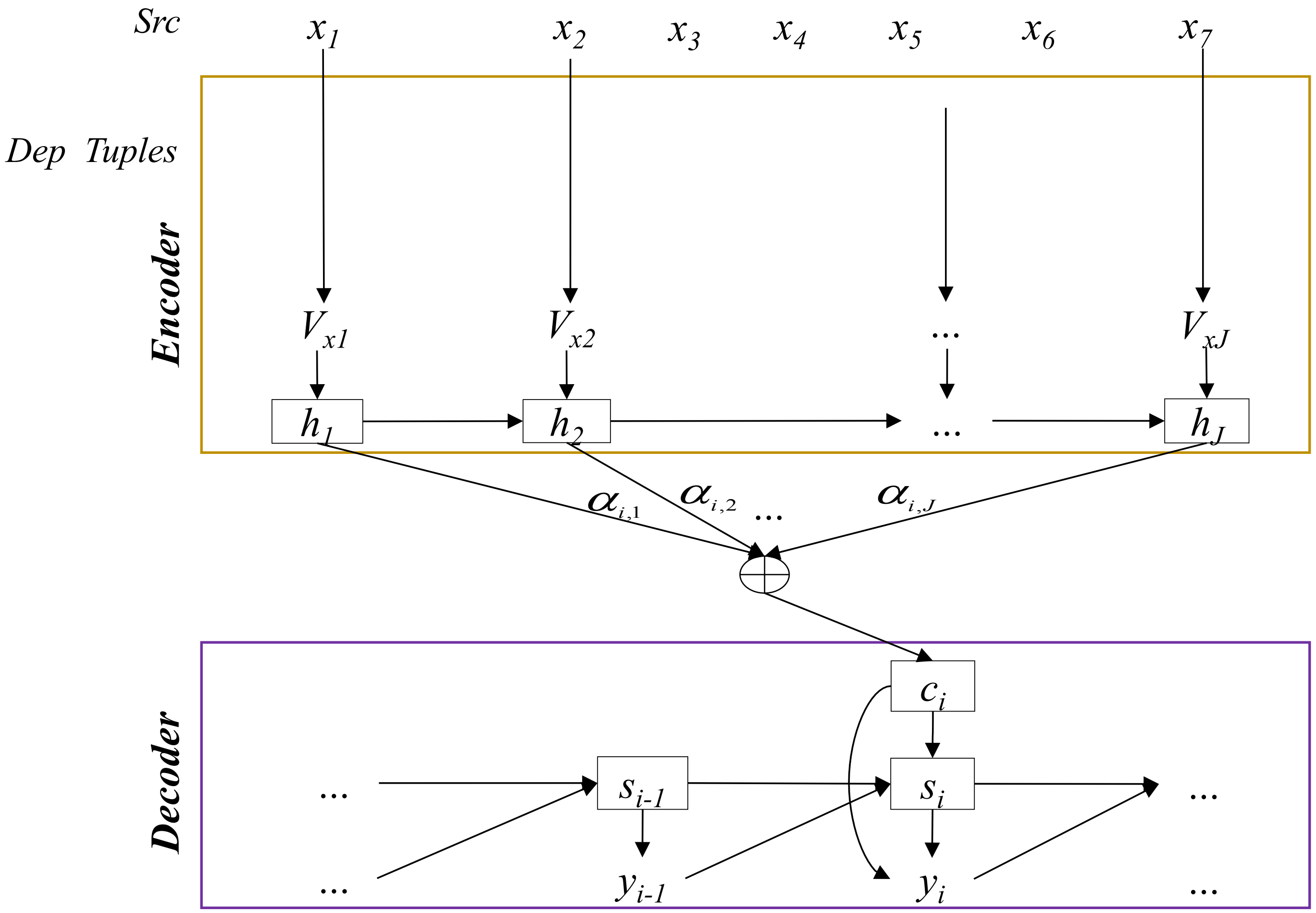
$SI_{x_2} =< x_1, x_4, x_7 >,$

$CH_{x_2} =< \varepsilon >,$



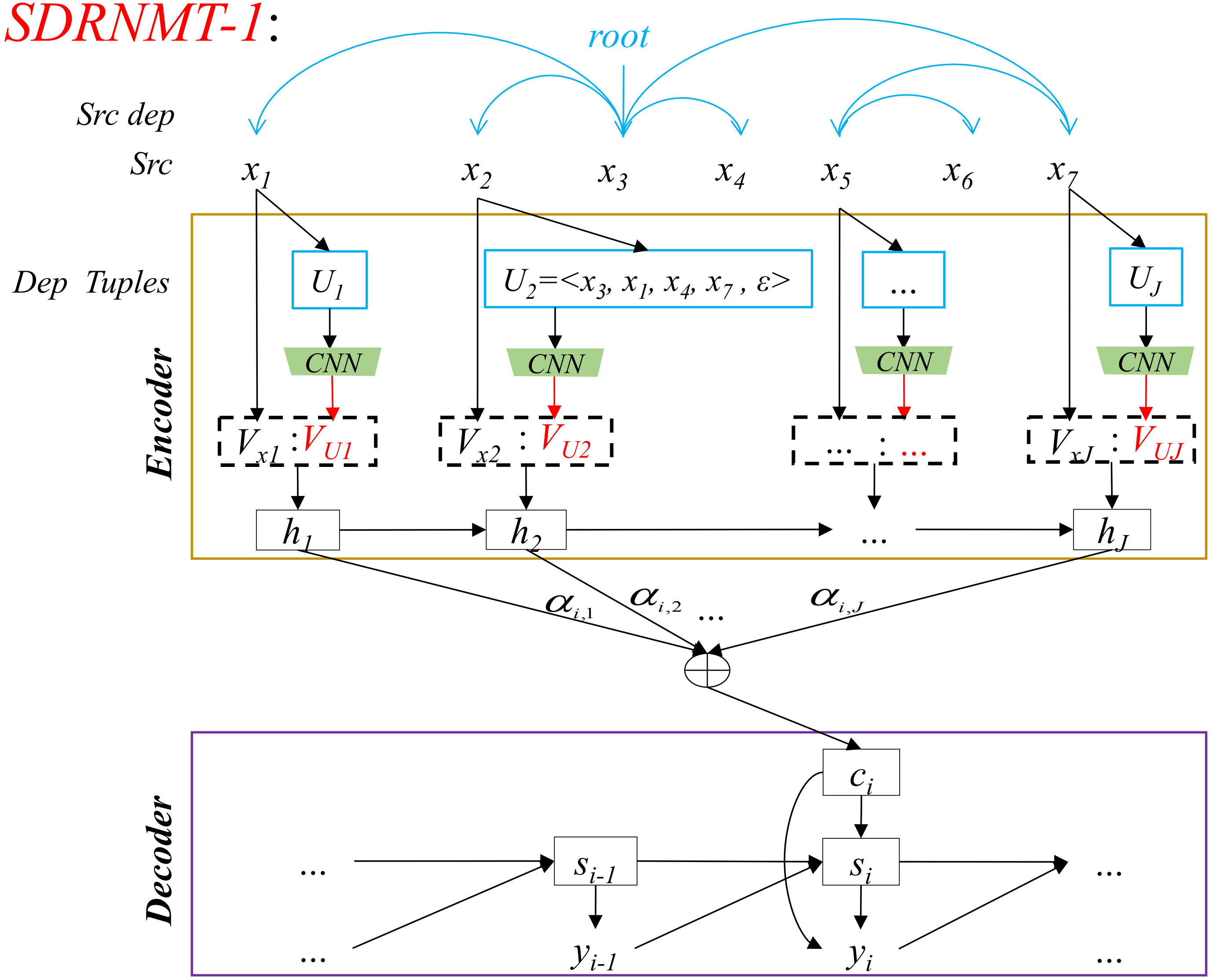
Neural Machine Translation with SDR

SDRNMT-1:



Neural Machine Translation with SDR

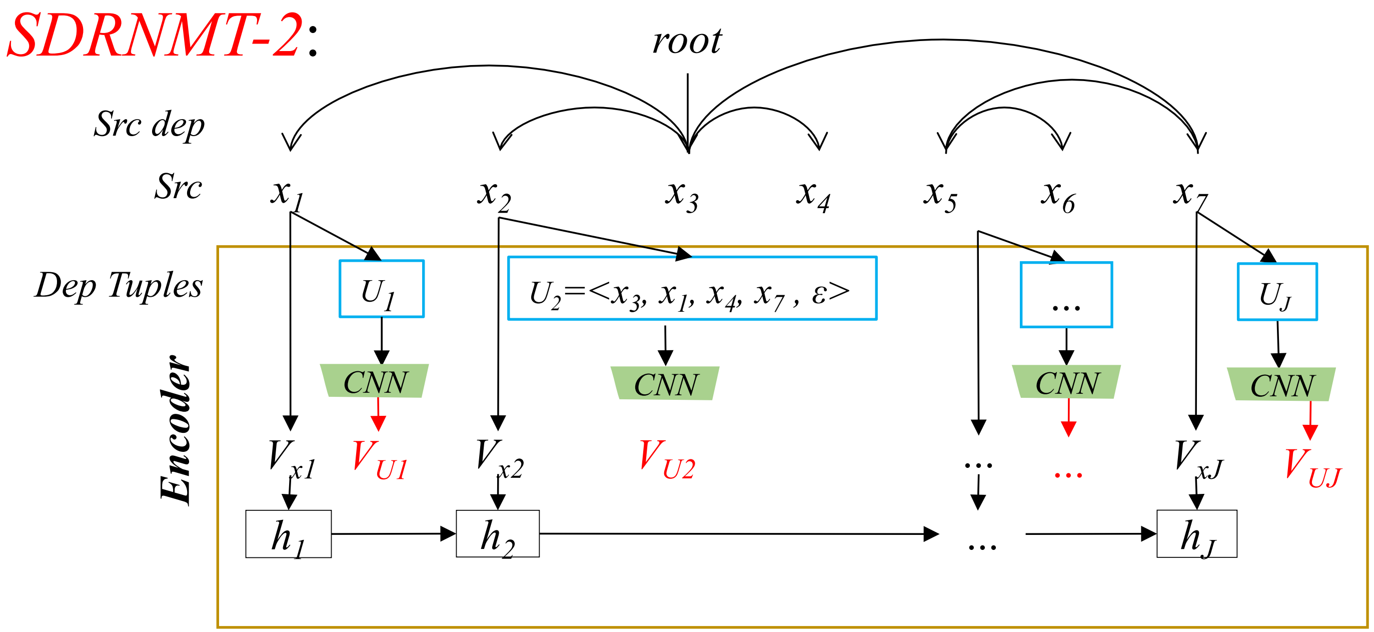
SDRNMT-1:



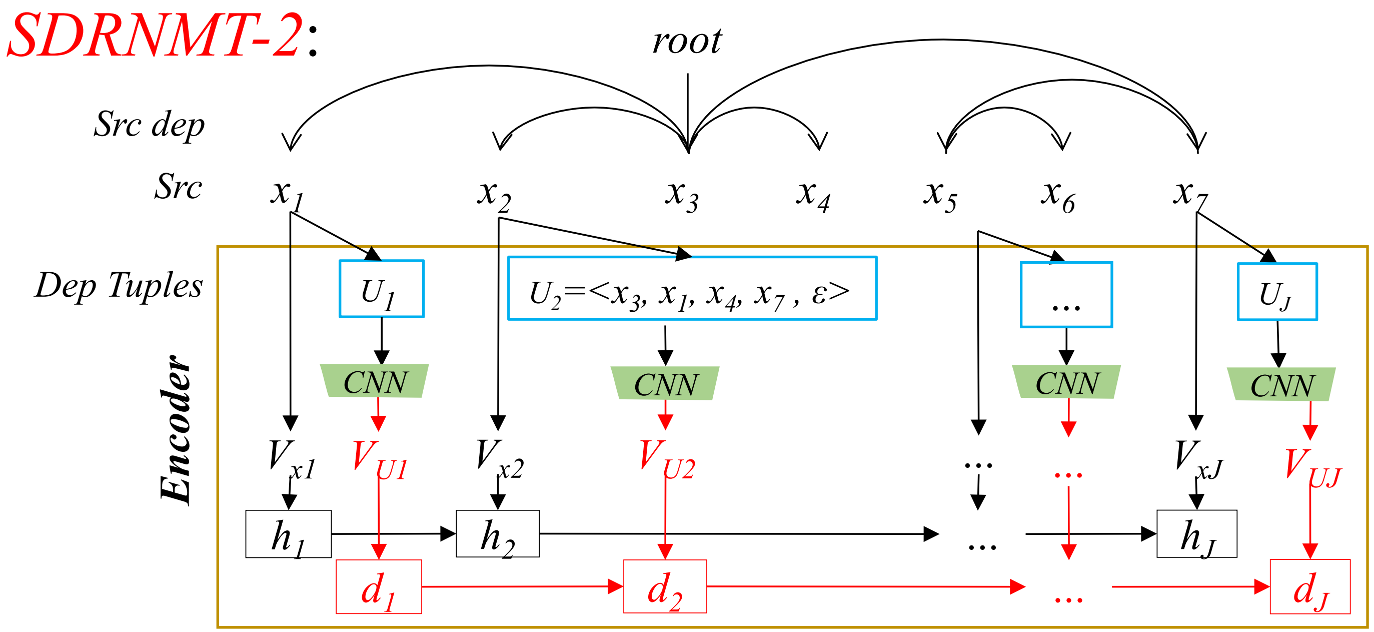
$$\boxed{h_j = f_{enc}(V_{x_j} : V_{U_j}, h_{j-1})}$$

Where the V_{xj} is 360-dim and the learned V_{Uj} is 260-dim.

Neural Machine Translation with SDR



Neural Machine Translation with SDR

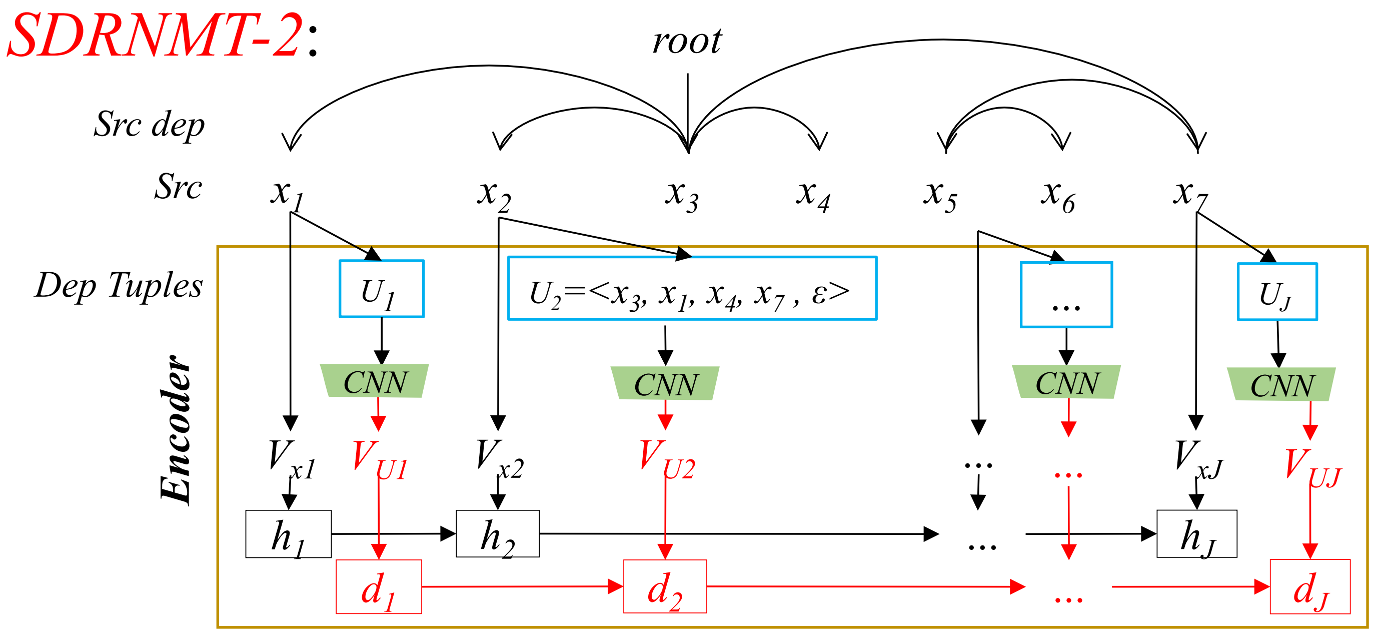


***Encoder*:**

$$h_j = f_{enc}(V_{x_j}, h_{j-1}),$$

$$d_j = f_{enc}(V_{U_j}, d_{j-1})$$

Neural Machine Translation with SDR



Attention $\tilde{\alpha}$

Encoder:

$$h_j = f_{enc}(V_{x_j}, h_{j-1}),$$

$$d_j = f_{enc}(V_{U_j}, d_{j-1})$$

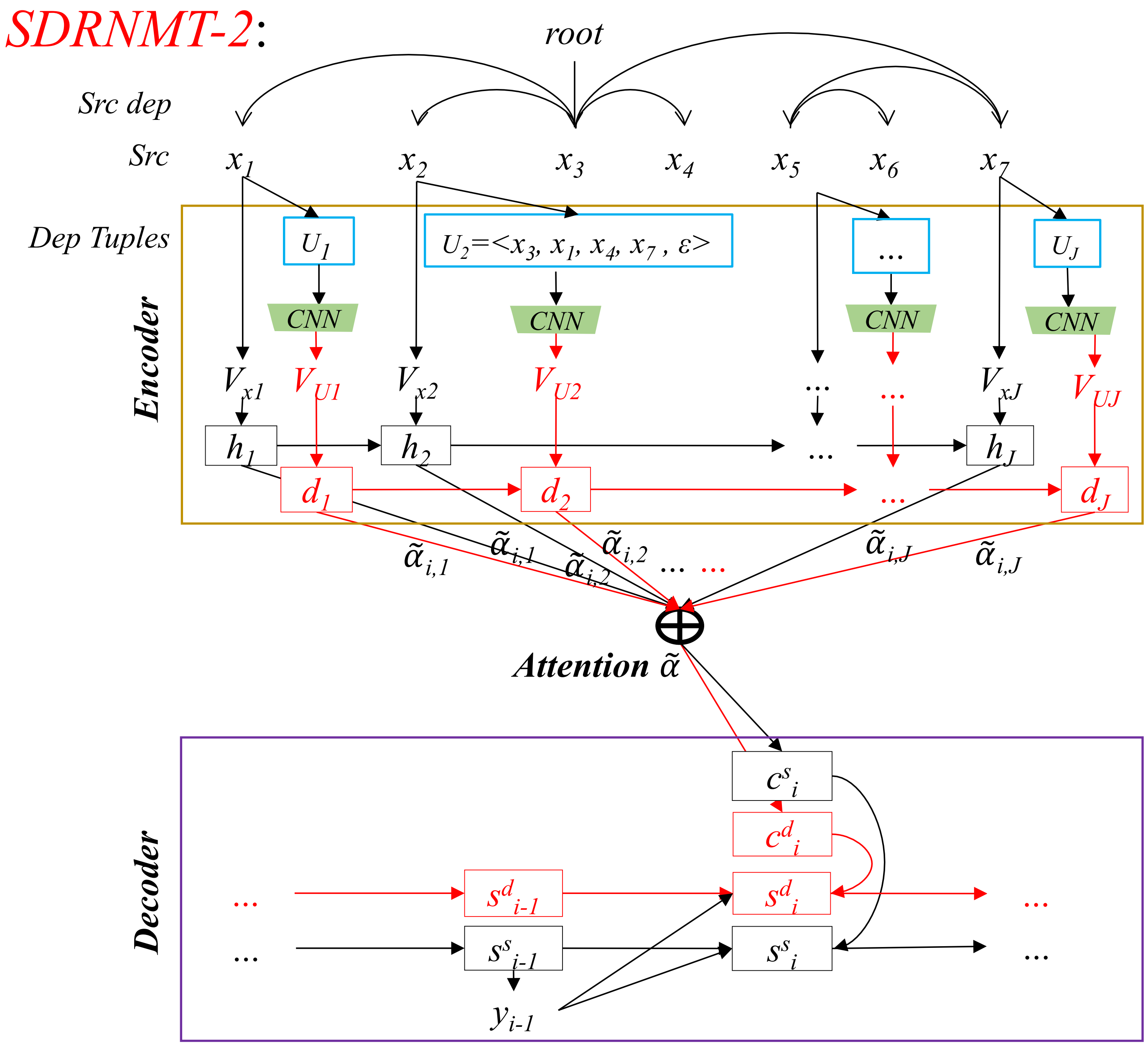
Attention:

$$e^s_{i,j} = f(s^s_{i-1} + h_j),$$

$$e^d_{i,j} = f(s^d_{i-1} + d_j).$$

$$\alpha_{i,j} = \frac{\exp(\lambda e^s_{i,j} + (1-\lambda) e^d_{i,j})}{\sum_{j=1}^{J} \exp(\lambda e^s_{i,j} + (1-\lambda) e^d_{i,j})}$$

Neural Machine Translation with SDR



Encoder:

$$h_j = f_{enc}(V_{x_j}, h_{j-1}),$$

$$d_j = f_{enc}(V_{U_j}, d_{j-1})$$

Attention:

$$e^s_{i,j} = f(s^s_{i-1} + h_j),$$

$$e^d_{i,j} = f(s^d_{i-1} + d_j).$$

$$\alpha_{i,j} = \frac{\exp(\lambda e^s_{i,j} + (1-\lambda) e^d_{i,j})}{\sum_{j=1}^{J} \exp(\lambda e^s_{i,j} + (1-\lambda) e^d_{i,j})}$$

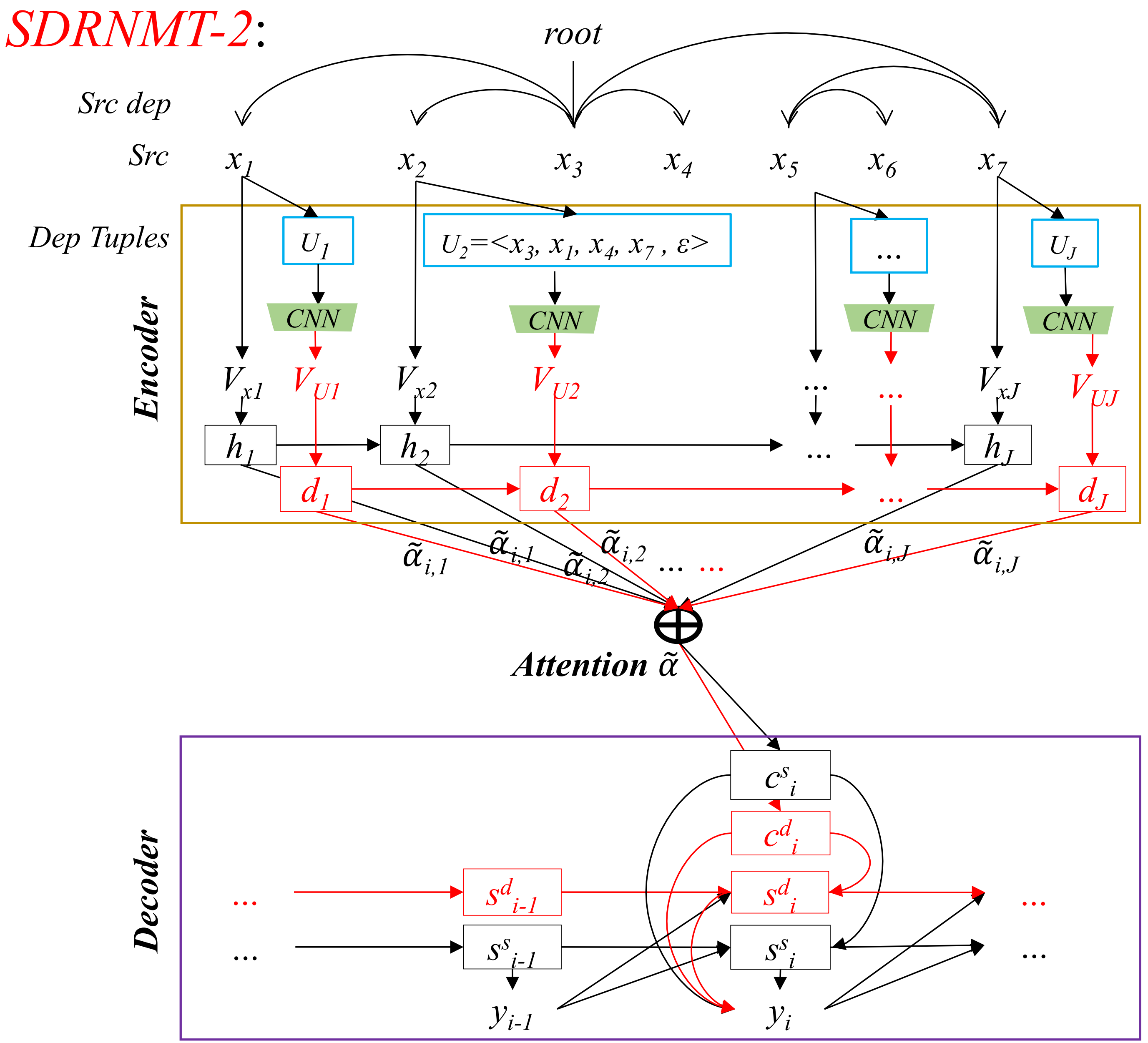
Decoder:

$$c^s_{i,j} = \sum_{j=1}^{J} \alpha_{i,j} h_j, c^d_{i,j} = \sum_{j=1}^{J} \alpha_{i,j} d_j$$

$$s^s_i = \varphi(s^s_{i-1}, y_{i-1}, c^s_i),$$

$$s^d_i = \varphi(s^d_{i-1}, y_{i-1}, c^d_i).$$

Neural Machine Translation with SDR



Encoder:
$$h_j = f_{enc}(V_{x_j}, h_{j-1}),$$
$$d_j = f_{enc}(V_{U_j}, d_{j-1})$$

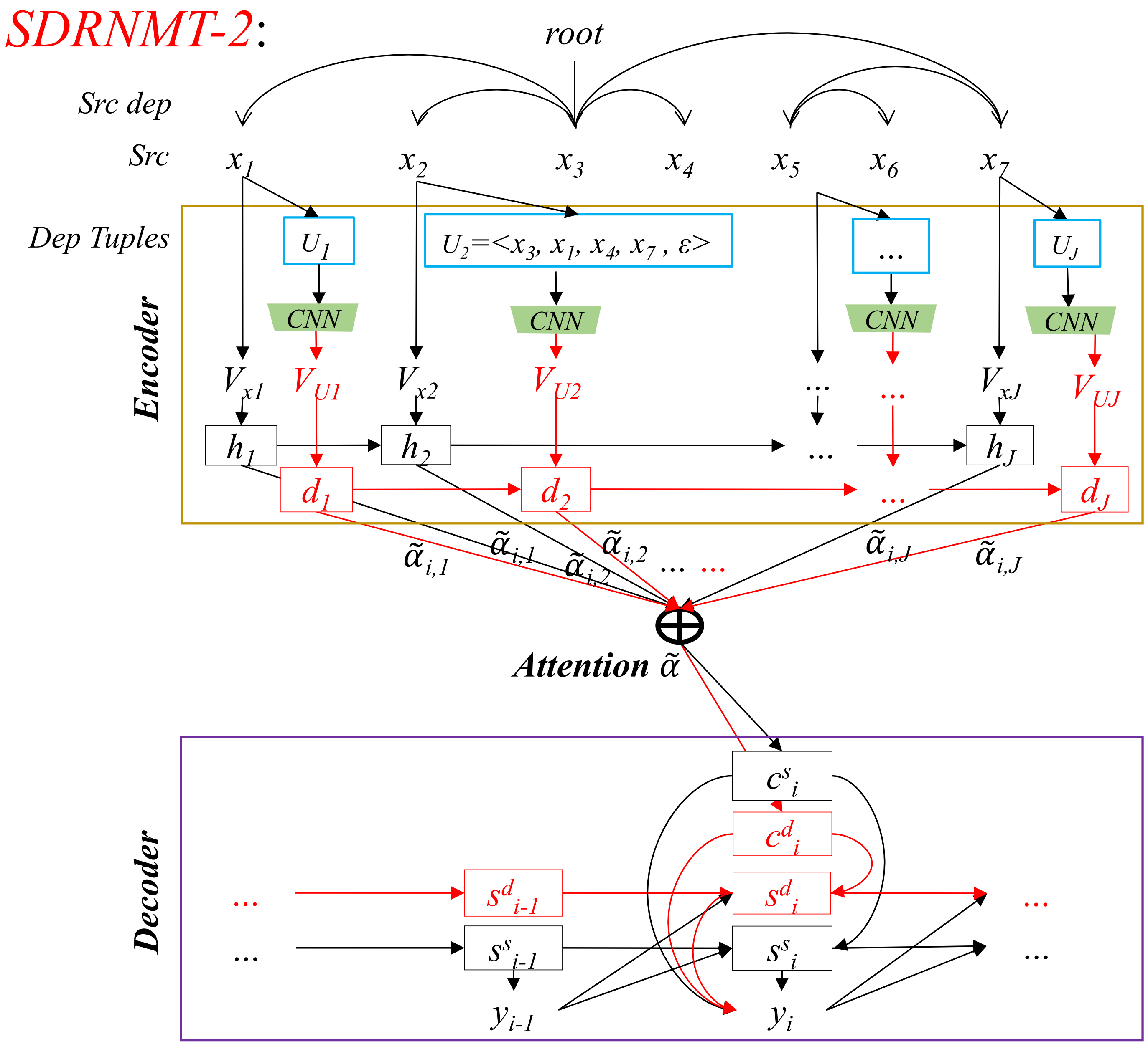
Attention:
$$e^s_{i,j} = f(s^s_{i-1} + h_j),$$
$$e^d_{i,j} = f(s^d_{i-1} + d_j).$$

$$\alpha_{i,j} = \frac{\exp(\lambda e^s_{i,j} + (1-\lambda) e^d_{i,j})}{\sum_{j=1}^{J} \exp(\lambda e^s_{i,j} + (1-\lambda) e^d_{i,j})}$$

Decoder:
$$c^s_{i,j} = \sum_{j=1}^{J} \alpha_{i,j} h_j, c^d_{i,j} = \sum_{j=1}^{J} \alpha_{i,j} d_j$$
$$s^s_i = \varphi(s^s_{i-1}, y_{i-1}, c^s_i),$$
$$s^d_i = \varphi(s^d_{i-1}, y_{i-1}, c^d_i).$$

$$p(y_i \mid y_{-i}, x, T) = g(y_{i-1}, s^s_i, s^d_i, c^s_i, c^d_i)$$

Neural Machine Translation with SDR



Encoder:

$$h_j = f_{enc}(V_{x_j}, h_{j-1}),$$

$$d_j = f_{enc}(V_{U_j}, d_{j-1})$$

Attention:

$$e^s_{i,j} = f(s^s_{i-1} + h_j),$$

$$e^d_{i,j} = f(s^d_{i-1} + d_j).$$

$$\alpha_{i,j} = \frac{\exp(\lambda e^s_{i,j} + (1-\lambda)e^d_{i,j})}{\sum_{j=1}^{J} \exp(\lambda e^s_{i,j} + (1-\lambda)e^d_{i,j})}$$

Decoder:

$$c^s_{i,j} = \sum_{j=1}^{J} \alpha_{i,j} h_j, c^d_{i,j} = \sum_{j=1}^{J} \alpha_{i,j} d_j$$

$$s^s_i = \varphi(s^s_{i-1}, y_{i-1}, c^s_i),$$

$$s^d_i = \varphi(s^d_{i-1}, y_{i-1}, c^d_i).$$

$$p(y_i \mid y_{-i}, x, T) = g(y_{i-1}, s^s_i, s^d_i, c^s_i, c^d_i)$$

Double Context NMT

Experimental

- Experiments on Chinese-to-English translation task, 1.42M *LDC corpus*
- Parse source sentences of training data by Stanford Parser (Chang et al.,2009)
- For the *SDRNMT-1* and *SDRNMT-2*, the dimension of V_{xj} is 360 and the dimension of V_{Uj} is 260, and input embedding of the baseline is 620
- The baselines include Phrase-Based Statistical Machine Translation (PBSMT) (Koehn et al., 2007), standard Attentional NMT (AttNMT) (Bahdanau et al., 2014), NMT with dependency labels (Sennrich and Haddow, 2016)

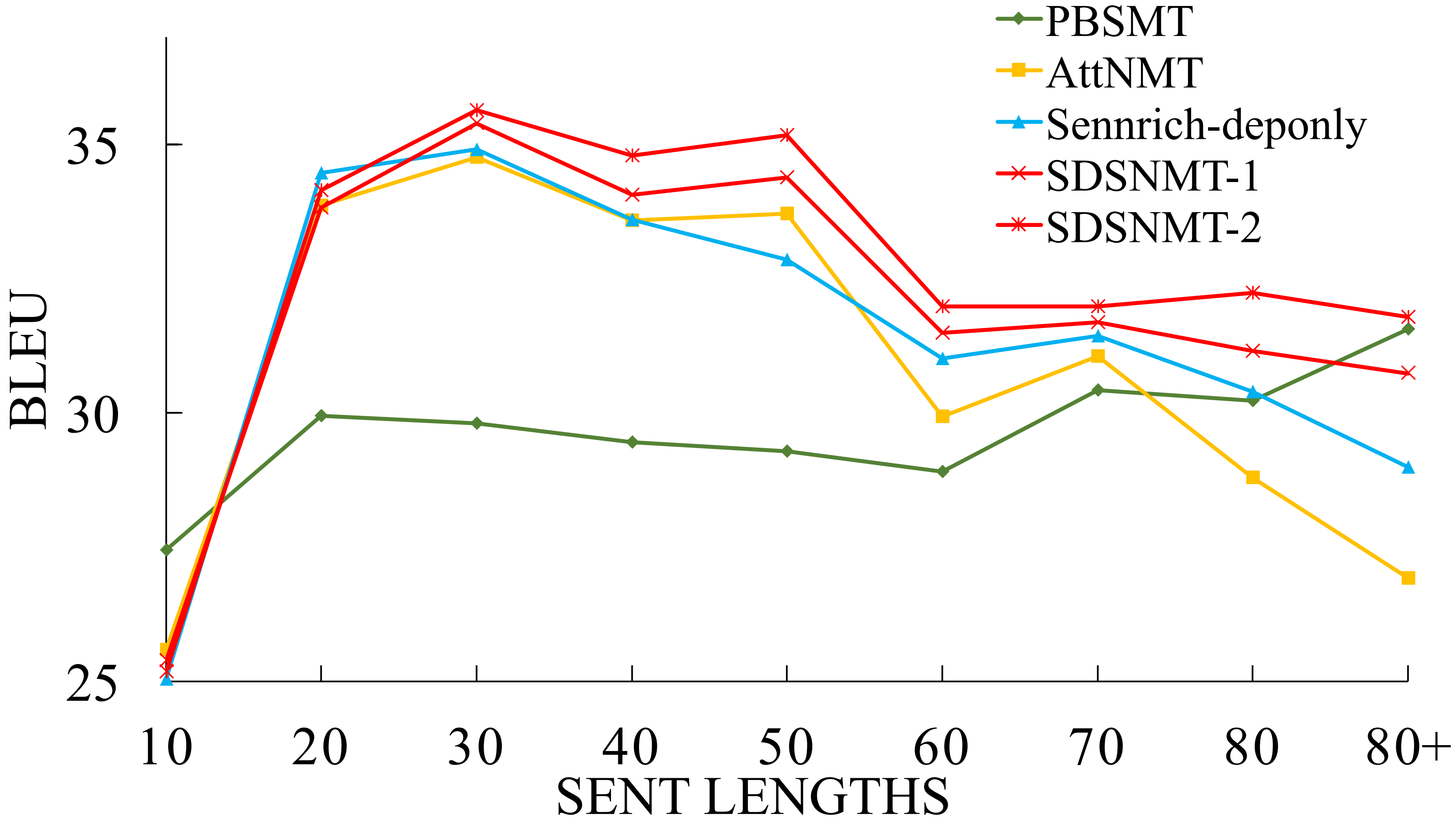
Experimental

System	Dev(NIST02)	NIST03	NIST04	NIST05	NIST06	NIST08	AVG
PBSMT	33.15	31.02	33.78	30.33	29.62	23.53	29.66
AttNMT	36.31	34.02	37.11	32.86	32.54	25.44	32.40
Sennrich-deponly	36.68	34.51	38.09	33.37	32.96	26.96	32.98
SDRNMT-*1*	36.88	34.98*	38.14	34.61**	33.58*	27.06	33.32
SDRNMT-*2*	**37.34**	**35.91****	**38.73***	**34.18***	**33.76****	**27.64***	**34.04**

"*" indicates statistically significant better than "Sennrich-deponly" at p-value < 0.05 and "**" at p-value < 0.01 by bootsrap resampling (Koehn, 2004)

Experimental Results

- Translation qualities for different sentence lengths



Conclusion

- Source dependency unit to capture source long-distance dependency constraint
- The proposed *SDRNMT-1* and *SDRNMT-2* consist of NMT and CNN, which are jointly trained to learn SDR and translation instead of separately trained
- Double-Context approach to further utilize source dependency representation