Neural Machine Translation with Source Dependency Representation

Kehai Chen\textsuperscript{1}, Rui Wang\textsuperscript{2}, Masao Utiyama\textsuperscript{2}, Lemao Liu\textsuperscript{3}, Akihiro Tamura\textsuperscript{4}, Eiichiro Sumita\textsuperscript{2} and Tiejun Zhao\textsuperscript{1}

\textsuperscript{1}Harbin Institute of Technology, China
\textsuperscript{2}National Institute of Information and Communications Technology, Japan
\textsuperscript{3}Tencent AI Lab, China
\textsuperscript{4}Ehime University, Japan
Overview

• Traditional NMT Model

\[ \text{Src: } x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7 \]

\[ \downarrow \]

Standard NMT model

\[ \text{Trg: } y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5 \quad y_6 \quad y_7 \quad y_8 \]
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:

$$U_j = \langle PA_{x_j}, SI_{x_j}, CH_{x_j} \rangle$$
Source Dependency Representation (SDR)

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

$$U_j = \langle PA_{x_j}, SI_{x_j}, CH_{x_j} \rangle$$

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

Take $x_2$ as an example:

$$PA_{x_2} = \langle x_3 \rangle,$$

then,

$$U_2 = \langle x_3, x_1, x_4, x_7, \varepsilon \rangle$$

$$SI_{x_2} = \langle x_1, x_4, x_7 \rangle,$$

$$CH_{x_2} = \langle \varepsilon \rangle,$$
Source Dependency Representation (SDR)

- Learn semantic representation of each dependency unit

  Take $x_2$ as an example: $\text{PA}_{x_2} = < x_3 >$, then, $U_2 = < x_3, x_1, x_4, x_7, e >$

  $\text{SI}_{x_2} = < x_1, x_4, x_7 >$, 

  $\text{CH}_{x_2} = < e >$. 

---

**Diagram:**

- Input layer: $10 \times d$
- Convolution layer 1: $3 \times d$ kernel
- Max-pooling layer 1: $8 \times d$
- Convolution layer 2: $4 \times d$
- Max-pooling layer 2: $2 \times d$
- Output layer: $1 \times d$
Neural Machine Translation with SDR

**SDRNMT-1:**

**Src**

**Dep Tuples**

**Encoder**

$$V_{x_1} \rightarrow h_1 \rightarrow V_{x_2} \rightarrow h_2 \rightarrow \ldots \rightarrow V_{x_J} \rightarrow h_J$$

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5 \rightarrow x_6 \rightarrow x_7$$

**Decoder**

$$s_{i-1} \rightarrow s_i \rightarrow y_i$$

$$\alpha_{i,2} \rightarrow \alpha_{i,3} \rightarrow \ldots \rightarrow \alpha_{i,J}$$

$$c_i$$

$$y_{i-1}$$

$$y_i$$
Neural Machine Translation with SDR

SDRNMT-1:

Where the $V_{xj}$ is 360-dim and the learned $V_{Uj}$ is 260-dim.
Neural Machine Translation with SDR

SDRNMT-2:

Src dep
Src
Dep Tuples

Encoder

$U_1$
$U_2 = \langle x_3, x_1, x_5, x_7, \epsilon \rangle$
$U_j$

$V_{x1}$
$V_{x2}$
$V_{xJ}$

$V_{U1}$
$V_{U2}$
$V_{UJ}$

$h_1$
$h_2$
$h_J$

root

$\times 1$ $\times 2$ $\times 3$ $\times 4$ $\times 5$ $\times 6$ $\times 7$
Neural Machine Translation with SDR

SDRNMT-2:

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

SDRNMT-2:

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

Attention: \[ e_{i,j}^s = f(s_{i-1}^s + h_j), \]
\[ e_{i,j}^d = f(s_{i-1}^d + d_j). \]
\[ \alpha_{i,j} = \frac{\exp(\lambda e_{i,j}^s + (1-\lambda)e_{i,j}^d)}{\sum_{j=1}^J \exp(\lambda e_{i,j}^s + (1-\lambda)e_{i,j}^d)} \]
Neural Machine Translation with SDR

SDRNMT-2:

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

Attention: \( e_{i,j}^s = f(s_{i-1}^s + h_j) \),
\( e_{i,j}^d = f(s_{i-1}^d + d_j) \).

\[ \alpha_{i,j} = \frac{\exp(\lambda e_{i,j}^s + (1-\lambda)e_{i,j}^d)}{\sum_{j=1}^J \exp(\lambda e_{i,j}^s + (1-\lambda)e_{i,j}^d)} \]

Decoder: \( c_{i,j}^s = \sum_{j=1}^J \alpha_{i,j} h_j, c_{i,j}^d = \sum_{j=1}^J \alpha_{i,j} d_j \)
\( s_{i}^s = \varphi(s_{i-1}^s, y_{i-1}, c_{i}^s) \),
\( s_{i}^d = \varphi(s_{i-1}^d, y_{i-1}, c_{i}^d) \).
Neural Machine Translation with SDR

**SDRNMT-2:**

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

**Attention:**
\[ e_{i,j}^s = f(s_{i-1}^s + h_j), \]
\[ e_{i,j}^d = f(s_{i-1}^d + d_j). \]

\[ \alpha_{i,j} = \frac{\exp(\lambda e_{i,j}^s + (1-\lambda)e_{i,j}^d)}{\sum_{j=1}^J \exp(\lambda e_{i,j}^s + (1-\lambda)e_{i,j}^d)} \]

**Decoder:**
\[ c_{i,j}^s = \sum_{j=1}^J \alpha_{i,j} h_j, c_{i,j}^d = \sum_{j=1}^J \alpha_{i,j} d_j \]
\[ s_{i-1}^s = \varphi(s_{i-1}^s, y_{i-1}, c_{i,j}^s), \]
\[ s_{i-1}^d = \varphi(s_{i-1}^d, y_{i-1}, c_{i,j}^d). \]

\[ p(y_i | y_{i-1}, x, T) = g(y_{i-1}, s_{i-1}^s, s_{i-1}^d, c_{i,j}^s, c_{i,j}^d) \]
Neural Machine Translation with SDR

SDRNMT-2:

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

Attention:
\[ e_{i,j}^s = f(s_{i-1}^s + h_j), \]
\[ e_{i,j}^d = f(s_{i-1}^d + d_j). \]

\[ \alpha_{i,j} = \frac{\exp(\lambda e_{i,j}^s + (1 - \lambda)e_{i,j}^d)}{\sum_{j=1}^{J} \exp(\lambda e_{i,j}^s + (1 - \lambda)e_{i,j}^d)} \]

Decoder:
\[ c_{i,j}^s = \sum_{j=1}^{J} \alpha_{i,j} h_j, c_{i,j}^d = \sum_{j=1}^{J} \alpha_{i,j} d_j \]
\[ s_{i}^s = \varphi(s_{i-1}^s, y_{i-1}, c_{i}^s), \]
\[ s_{i}^d = \varphi(s_{i-1}^d, y_{i-1}, c_{i}^d). \]
\[ p(y_i | y_{i-1}, x, T) = g(y_{i-1}, s_{i}^s, s_{i}^d, c_{i}^s, c_{i}^d) \]

Double Context NMT
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

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

“*” indicates statistically significant better than “Sennrich-deponly” at $p$-value $< 0.05$ and “**” at $p$-value $< 0.01$ by bootstrap 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