Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences

Hongyuan Mei, Mohit Bansal, Matthew R. Walter
Toyota Technological Institute, Chicago
Introduction

• Neural sequence-to-sequence model for direction following
Place your back against the wall of the “T” intersection. Go forward one segment to the intersection with the blue-tiled hall. This intersection [sic] contains a chair. Turn left. Go forward to the end of the hall. Turn left. Go forward one segment to the intersection with the wooden-floored hall. This intersection contains [sic] an easel. Turn right. Go forward two segments to the end of the hall. Turn left. Go forward one segment to the intersection containing the lamp. Turn right. Go forward one segment to the empty corner.
Introduction

- Learn correspondences between instruction and actions using an alignment-based LSTM
- End-to-end differentiable sequence-to-sequence model
Model architecture

Figure 2: Our encoder-aligner-decoder model with multi-level alignment
Model architecture

• Inference over a probabilistic model

\[ a^*_1:T = \arg \max_{a_1:T} P(a_1:T|y_1:T, x_1:N) \]

\[ = \arg \max_{a_1:T} \prod_{t=1}^{T} P(a_t|a_{1:t-1}, y_t, x_1:N) \]

• Neural encoder decoder model with attention
Model architecture

- Bidirectional LSTM to encode instruction

\[
\begin{pmatrix}
    i_j^e \\
    f_j^e \\
    o_j^e \\
    g_j^e
\end{pmatrix} =
\begin{pmatrix}
    \sigma \\
    \sigma \\
    \sigma \\
    \tanh
\end{pmatrix} T_j^e \begin{pmatrix}
    x_j \\
    h_{j-1}
\end{pmatrix}
\]

\[
c_j^e = f_j^e \odot c_{j-1}^e + i_j^e \odot g_j^e
\]

\[
h_j = o_j^e \odot \tanh(c_j^e)
\]

\[
h_j = (\overrightarrow{h_j}, \overleftarrow{h_j})^T
\]
Model architecture

• Multi level aligner: High level (hidden states of LSTM) + low level (input words)
• One layer neural perceptron

\[ z_t = \sum_j \alpha_{tj} \left( x_j \right) \]
\[ \alpha_{tj} = \exp(\beta_{tj}) / \sum_j \exp(\beta_{tj}), \]
\[ \beta_{tj} = v^T \tanh(W s_{t-1} + U x_j + V h_j) \]

• Intuitively, better match the salient words in input sentence (e.g., “easel”) directly to corresponding landmarks in the current world state \( \hat{y}(t) \) used in decoder
Model architecture

• LSTM decoder

\[
\begin{pmatrix}
    i_t^d \\
    f_t^d \\
    o_t^d \\
    g_t^d
\end{pmatrix} =
\begin{pmatrix}
    \sigma \\
    \sigma \\
    \sigma \\
    \text{tanh}
\end{pmatrix} T^d
\begin{pmatrix}
    Ey_t \\
    s_{t-1} \\
    z_t
\end{pmatrix}
\]

\[
c_t^d = f_t^d \odot c_{t-1}^d + i_t^d \odot g_t^d
\]

\[
s_t = o_t^d \odot \text{tanh}(c_t^d)
\]

\[
q_t = L_0(Ey_t + L_s s_t + L_z z_t)
\]

\[
P_{a,t} = \text{softmax} \ (q_t)
\]

• Output P is the conditional probability distribution over actions
• E is an embedding matrix
• Trained using negative log likelihood of demonstrated action
Experiments

• SAIL route instructor dataset
• World state ($y(t)$) encodes local observable world at time $t$, encoded as a concatenation of a bag-of-words vector for each direction (forward, left, and right).
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Single-sent</th>
<th>Multi-sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen and Mooney (2011)</td>
<td>54.40</td>
<td>16.18</td>
</tr>
<tr>
<td>Chen (2012)</td>
<td>57.28</td>
<td>19.18</td>
</tr>
<tr>
<td>Kim and Mooney (2012)</td>
<td>57.22</td>
<td>20.17</td>
</tr>
<tr>
<td>Kim and Mooney (2013)</td>
<td>62.81</td>
<td>26.57</td>
</tr>
<tr>
<td>Artzi and Zettlemoyer (2013)</td>
<td>65.28</td>
<td>31.93</td>
</tr>
<tr>
<td>Artzi, Das, and Petrov (2014)</td>
<td>64.36</td>
<td><strong>35.44</strong></td>
</tr>
<tr>
<td>Andreas and Klein (2015)</td>
<td>59.60</td>
<td>–</td>
</tr>
<tr>
<td>Our model (vDev)</td>
<td>69.98</td>
<td>26.07</td>
</tr>
<tr>
<td>Our model (vTest)</td>
<td><strong>71.05</strong></td>
<td>30.34</td>
</tr>
</tbody>
</table>
## Ablation results

<table>
<thead>
<tr>
<th></th>
<th>Full Model</th>
<th>High-level Aligner</th>
<th>No Aligner</th>
<th>Unidirectional</th>
<th>No Encoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-sentence</td>
<td>69.98</td>
<td>68.09</td>
<td>68.05</td>
<td>67.44</td>
<td>61.63</td>
</tr>
<tr>
<td>Multi-sentence</td>
<td>26.07</td>
<td>24.79</td>
<td>25.04</td>
<td>24.50</td>
<td>16.67</td>
</tr>
</tbody>
</table>

Table 2: Model components ablations
Visualization

Figure 4: Visualization of the alignment between words to actions in a map for a multi-sentence instruction.