JBNU at MRP 2019: Multi-level Biaffine Attention for Semantic Dependency Parsing

Seung-Hoon Na, Jinwoo Min, Kwanghyeon Park
Dept. Computer Science, Jeonbuk National University

Jong-Hun Shin, Young-Kil Kim
Electronics and Telecommunication Research Institute (ETRI)
Introduction

• Our issue: **Multi-task learning** for **DM/PSD/UCCA**
  – To enable **multi-task learning**, we explicitly make **shared common components** in a neural network architecture across different frameworks

• Models
  – **Biaffine attention**: we propose a unified neural model for the **DM/PSD/UCCA** frameworks based on the biaffine attention [Dozat and Manning, 2017, 2018; Zhang et al., 2019]
  – **Multi-level biaffine attention**:
    • Motivated by the multi-level architecture of FusionNet in the machine reading comprehension task [Huang et al., 2018]
This technique is impossible to apply…

**Encoder:** BERT-BiLSTM ← shared across frameworks

**Decoder:** Biaffine attention ← framework specific

Word representation layer

---

DM → PSD → UCCA
Encoder: BERT-BiLSTM (shared across frameworks)

- **Word representation layer using BERT**

  Given a sentence, the BERT encoder is applied to its wordpieces and the encoded wordpiece-level representations are composed to the word-level embeddings based on BiLSTM.

\[
x_i = [w_i^{bert}; e_i^{glove}; e_i^{POS}]
\]

\[
r_i = BiLSTM_i(x_1 \cdots x_n)
\]
Decoder: Biaffine attention (framework specific)

- Biaffine attention is performed on the role-dependent representations to predict the existence of an edge and its labels.

Biaffine attention

\[ BiAff_{m}(x, y) = x^T U^{[1:m]} y + V \begin{bmatrix} x \\ y \end{bmatrix} + b \]

\[ s_{i,j}^{(edge)} = BiAff_{1}^{(edge)} \left( h_{i}^{(dep)}, h_{j}^{(head)} \right) \]

\[ s_{i,j}^{(label)} = BiAff_{k}^{(label)} \left( h_{i}^{(l-dep)}, h_{j}^{(l-head)} \right) \]

\[ s_{i}^{(top)} = FFN^{(top)} (r_{i}) \]
Multi-level Biaffine attention

The hidden representations at three levels are composed to the final hidden representation $z_i^{(dep)}, z_i^{(head)}$ using a semantic fusion unit
Preliminary Experiment

For more details, please visit our poster. Thank you.

<table>
<thead>
<tr>
<th>method</th>
<th>DM</th>
<th></th>
<th>PSD</th>
<th></th>
<th>UCCA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top</td>
<td>UF</td>
<td>LF</td>
<td>Top</td>
<td>UF</td>
<td>LF</td>
</tr>
<tr>
<td>Biaffine</td>
<td>93.67</td>
<td>92.08</td>
<td>90.86</td>
<td>95.97</td>
<td>90.50</td>
<td>78.21</td>
</tr>
<tr>
<td>BERT+Biaffine</td>
<td>95.06</td>
<td>93.85</td>
<td>93.00</td>
<td>96.89</td>
<td>92.30</td>
<td>80.24</td>
</tr>
<tr>
<td>BERT+Multi-level Biaffine</td>
<td>95.09</td>
<td>93.86</td>
<td>93.02</td>
<td>96.76</td>
<td>91.95</td>
<td>79.76</td>
</tr>
<tr>
<td>BERT+Biaffine+MTL</td>
<td>N/A</td>
<td>93.66</td>
<td>92.73</td>
<td>N/A</td>
<td>92.13</td>
<td>79.63</td>
</tr>
</tbody>
</table>

- **BERT+Biaffine performs** better than Biaffine, in particular, obtaining the increases of about 5% for UF and LF on the UCCA framework
- **BERT+Multi-level Biaffine** does not achieve any further improvements with respect to BERT-Biaffine model
- **BERT+Biaffine+MTL** only achieves small improvements on UCCA framework whereas no improvements on DM and PSD frameworks can be observed.

<table>
<thead>
<tr>
<th>Framework</th>
<th>Train</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>32091</td>
<td>3565</td>
</tr>
<tr>
<td>PSD</td>
<td>32091</td>
<td>3565</td>
</tr>
<tr>
<td>UCCA</td>
<td>5915</td>
<td>656</td>
</tr>
</tbody>
</table>