PCFGs Can Do Better: Inducing Probabilistic Context-Free Grammars with Many Symbols

Songlin Yang$^1$, Yanpeng Zhao$^2$, Kewei Tu$^1$

$^1$School of Information Science and Technology, ShanghaiTech University

$^2$ILCC, University of Edinburgh

Introduction

Neural PCFGs have been proved effective in unsupervised constituency parsing. However, due to the high complexity of the inside algorithm, they cannot scale to a relatively large number of grammar symbols. We use CP decomposition to decompose the binary rule probability tensor to decrease the complexity from cubic to at most quadratic in the number of grammar symbols. We enforce row/column normalization on the decomposed matrices to ensure the validity of the binary rule probability.

CP decomposition on PCFGs

Recursive form of the inside algorithm

$$s_{ij}^r = \sum_{k=1}^{j-1} \sum_{l \in B,C} \pi_{A \rightarrow BC} \cdot s_{ik}^p \cdot s_{kl+1}^c$$

(1)

Kruskal form of the binary rule probability tensor:

$$T = q_{ij}^{(1)} \cdot T^{(0)} = u^{(1)} \odot v^{(0)} \odot w^{(0)}$$

Eq. 1 becomes the following with complexity $O(pm + \beta mr)$:

$$s_{ij} = u \cdot \sum_{k=1}^{j-1} (V^T s_{ik} \odot (W^T s_{k+1}))$$

Neural Parameterization

Use distributed representation for each symbol. Use neural networks to calculate grammar rule probabilities. Use Softmax function to ensure the row/column normalization of $U, V, W$.

Without neural parameterization, performance drops significantly (even underperforms right-branching baseline in WSJ)

Activation functions other than ReLU perform much worse.

Experiments

Comparison on WSJ test set.

Influence of symbol number

Figure 3. We omit the PCFG suffix. TN: our proposed method. X: X preterminals. N: Neural. NL: Neural Lexicalized. C: Compound

Figure 4. We set the ratio of nonterminals: preterminals as 1:2.

Multilingual experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Chinese</th>
<th>Basque</th>
<th>German</th>
<th>French</th>
<th>Hebrew</th>
<th>Hungarian</th>
<th>Korean</th>
<th>Polish</th>
<th>Swedish</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left Branching$^1$</td>
<td>7.2</td>
<td>17.9</td>
<td>10.0</td>
<td>5.7</td>
<td>8.5</td>
<td>13.3</td>
<td>18.5</td>
<td>10.9</td>
<td>8.4</td>
<td>11.2</td>
</tr>
<tr>
<td>Right Branching$^1$</td>
<td>25.5</td>
<td>15.4</td>
<td>14.7</td>
<td>26.4</td>
<td>30.0</td>
<td>12.7</td>
<td>19.2</td>
<td>34.2</td>
<td>30.4</td>
<td>23.2</td>
</tr>
<tr>
<td>Random Trees$^1$</td>
<td>15.2</td>
<td>19.5</td>
<td>13.9</td>
<td>16.2</td>
<td>19.7</td>
<td>14.1</td>
<td>22.2</td>
<td>16.1</td>
<td>17.4</td>
<td>16.4</td>
</tr>
<tr>
<td>N-PCFG w/ MBR</td>
<td>26.3$^{\pm0.5}$</td>
<td>35.1$^{\pm0.4}$</td>
<td>42.8$^{\pm0.7}$</td>
<td>45.0$^{\pm0.2}$</td>
<td>45.7$^{\pm0.2}$</td>
<td>43.5$^{\pm0.4}$</td>
<td>38.4$^{\pm0.5}$</td>
<td>43.2$^{\pm0.3}$</td>
<td>41.0$^{\pm0.7}$</td>
<td>39.2$^{\pm0.6}$</td>
</tr>
<tr>
<td>C-PCFG w/ MBR</td>
<td>38.7$^{\pm0.3}$</td>
<td>40.4$^{\pm0.2}$</td>
<td>43.5$^{\pm0.6}$</td>
<td>45.0$^{\pm0.1}$</td>
<td>45.2$^{\pm0.4}$</td>
<td>44.9$^{\pm0.0}$</td>
<td>30.5$^{\pm0.8}$</td>
<td>43.8$^{\pm0.3}$</td>
<td>33.0$^{\pm0.4}$</td>
<td>40.1$^{\pm0.8}$</td>
</tr>
<tr>
<td>TN-PCFG $p=500$</td>
<td>39.2$^{\pm0.0}$</td>
<td>36.0$^{\pm1.0}$</td>
<td>47.1$^{\pm1.7}$</td>
<td>39.1$^{\pm1.1}$</td>
<td>39.2$^{\pm1.0}$</td>
<td>43.1$^{\pm1.1}$</td>
<td>35.4$^{\pm0.8}$</td>
<td>48.6$^{\pm1.1}$</td>
<td>40.0$^{\pm0.8}$</td>
<td>40.9$^{\pm0.8}$</td>
</tr>
</tbody>
</table>