Multilingual Grammar Induction with Continuous Language Identification

Wenjuan Han, Ge Wang, Yong Jiang, Kewe Tu

ShanghaiTech University, Shanghai, China
Alibaba Group
{hanwj,wangge,tukw}@shanghaitech.edu.cn
{yongjiang.jy}@alibaba-inc.com

November 9, 2019
Outline

1. Motivation
2. Model
3. Learning
4. Experiments
5. Conclusion
Outline

1 Motivation
2 Model
3 Learning
4 Experiments
5 Conclusion
Grammar induction is the task to learn grammars form unannotated corpus.
Grammar induction is the task to learn grammars from unannotated corpus.

Multilingual grammar induction couples grammar parameters of different languages together and learns them simultaneously.
• Grammar induction is the task to learn grammars from unannotated corpus.

• Multilingual grammar induction couples grammar parameters of different languages together and learns them simultaneously.

→ The key is to exploit the similarities between languages.
Existing approaches to tackle this problem:

- Treating languages equally (Iwata et al., 2010).
- Utilizing hand-crafted phylogenetic tree to encode this kind of information (Berg-Kirkpatrick and Klein, 2010).
Existing approaches to tackle this problem:

- Treating languages equally (Iwata et al., 2010). → Language similarity ignored.

- Utilizing hand-crafted phylogenetic tree to encode this kind of information (Berg-Kirkpatrick and Klein, 2010). → Need linguistic knowledge and sometimes could be misleading. Example: English is dominant SVO while German is not, although they are both Germanic languages.
Outline

1 Motivation

2 Model

3 Learning

4 Experiments

5 Conclusion
We represent language identities with continuous vectors i.e., language embeddings and use them to encode language similarity.
Model Architecture

Neural DMV grammar rule probability:
\[ P_{\text{ATTACH}}(\text{child}|\text{head}, \text{direction}, \text{valence}) \]
Neural DMV grammar rule probability:

\[ P_{\text{ATTACH}}(\text{child}|\text{head}, \text{direction}, \text{valence}) \]

Now we have: \( P_{\text{ATTACH}}(\text{child}|\text{head}, \text{direction}, \text{valence}, \text{language}) \)
Model Architecture

Predict language identification with language embeddings and sentence representations.

Han et al., 2019
Model Architecture

For each training sentence $x^{(i)}$ from language $l$:

- $P(x^{(i)}|G_{l(i)})$, the probability of the training sentence $x^{(i)}$ being generated from grammar $G_{l(i)}$.
- $P(l^{(i)}|x^{(i)})$, the probability of correct language identification of $x^{(i)}$. 

Han et al., 2019  
M-NDMV  
November 9, 2019 14 / 24
Outline

1. Motivation
2. Model
3. Learning
4. Experiments
5. Conclusion
Objective

For each training sentence $x^{(i)}$:

- $P(x^{(i)}|G_{l(i)})$, the probability of the training sentence $x^{(i)}$ being generated from grammar $G_{l(i)}$.
- $P(l^{(i)}|x^{(i)})$, the probability of correct language identification of $x^{(i)}$.

The training objective is:

$$\mathcal{L}(\Theta) = \sum_{(x, l) \in D} \left( \log P_\Theta(x|G_l) + \lambda \log P_\Theta(l|x) \right)$$
Learning

- $P(x^{(i)}|G^{(i)}) \rightarrow$ this term is optimized with EM (Adam used in M step).
- $P(I^{(i)}|x^{(i)}) \rightarrow$ Adam to optimize this term.
Outline

1 Motivation
2 Model
3 Learning
4 Experiments
5 Conclusion
## Dataset

We selected 15 languages across 8 language families and subfamilies from UD dataset to ensure diversity.

<table>
<thead>
<tr>
<th>Language</th>
<th>UD Treebank</th>
<th>Language Family</th>
<th>Corpus Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ET</td>
<td>Estonian</td>
<td>Finnic</td>
<td>11404</td>
</tr>
<tr>
<td>FI</td>
<td>Finnish</td>
<td>Finnic</td>
<td>9648</td>
</tr>
<tr>
<td>NL</td>
<td>Dutch</td>
<td>Germanic</td>
<td>8783</td>
</tr>
<tr>
<td>EN</td>
<td>English</td>
<td>Germanic</td>
<td>7674</td>
</tr>
<tr>
<td>DE</td>
<td>German</td>
<td>Germanic</td>
<td>7447</td>
</tr>
<tr>
<td>NO</td>
<td>Norwegian</td>
<td>Germanic</td>
<td>10017</td>
</tr>
<tr>
<td>GRC</td>
<td>Ancient_Greek</td>
<td>Hellenic</td>
<td>9387</td>
</tr>
<tr>
<td>HI</td>
<td>Hindi</td>
<td>Indo-Iranian</td>
<td>4997</td>
</tr>
<tr>
<td>JA</td>
<td>Japanese</td>
<td>Japonic</td>
<td>7441</td>
</tr>
<tr>
<td>FR</td>
<td>French</td>
<td>Romance</td>
<td>4976</td>
</tr>
<tr>
<td>IT</td>
<td>Italian</td>
<td>Romance</td>
<td>6492</td>
</tr>
<tr>
<td>LA</td>
<td>Latin-ITTB</td>
<td>Romance</td>
<td>10136</td>
</tr>
<tr>
<td>BG</td>
<td>Bulgarian</td>
<td>Slavonic</td>
<td>6507</td>
</tr>
<tr>
<td>SL</td>
<td>Slovenian</td>
<td>Slavonic</td>
<td>3800</td>
</tr>
<tr>
<td>EU</td>
<td>Basque</td>
<td>Vasconic</td>
<td>4271</td>
</tr>
</tbody>
</table>
Comparison of monolingual and multilingual approaches.

- **G**: our multilingual grammar model.
- **G+I**: our multilingual grammar model and auxiliary language identification task.

<table>
<thead>
<tr>
<th>Code</th>
<th>Monolingual</th>
<th>Multilingual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DMV NDMV</td>
<td>DMV NDMV G G+I</td>
</tr>
<tr>
<td>ET</td>
<td>51.8 52.9</td>
<td>43.1 45.3 56.0 56.4</td>
</tr>
<tr>
<td>FI</td>
<td>31.8 27.6</td>
<td>39.1 40.0 50.7 49.3</td>
</tr>
<tr>
<td>NL</td>
<td>42.4 35.6</td>
<td>46.5 47.8 50.4 50.6</td>
</tr>
<tr>
<td>EN</td>
<td>51.8 53.7</td>
<td>47.7 50.8 51.7 52.7</td>
</tr>
<tr>
<td>DE</td>
<td>52.8 50.4</td>
<td>55.5 57.2 59.6 61.4</td>
</tr>
<tr>
<td>NO</td>
<td>58.9 59.2</td>
<td>55.7 58.8 61.0 61.3</td>
</tr>
<tr>
<td>GRC</td>
<td>40.4 37.7</td>
<td>41.1 40.8 46.8 46.2</td>
</tr>
<tr>
<td>HI</td>
<td>52.6 53.9</td>
<td>29.2 31.1 47.4 46.8</td>
</tr>
<tr>
<td>JA</td>
<td>39.8 37.1</td>
<td>27.8 29.6 43.4 44.2</td>
</tr>
<tr>
<td>FR</td>
<td>58.8 38.1</td>
<td>59.6 59.4 58.4 60.1</td>
</tr>
<tr>
<td>IT</td>
<td>60.8 63.6</td>
<td>66.7 66.4 64.4 65.9</td>
</tr>
<tr>
<td>LA</td>
<td>32.6 36.3</td>
<td>39.8 42.0 45.1 45.0</td>
</tr>
<tr>
<td>BG</td>
<td>58.9 61.8</td>
<td>65.9 69.4 71.3 71.3</td>
</tr>
<tr>
<td>SL</td>
<td>70.7 67.5</td>
<td>62.1 63.3 68.3 68.6</td>
</tr>
<tr>
<td>EU</td>
<td>42.1 45.5</td>
<td>45.7 45.2 54.2 53.6</td>
</tr>
<tr>
<td>Avg</td>
<td>49.7 48.1</td>
<td>48.4 49.8 55.3 55.6</td>
</tr>
</tbody>
</table>

Each language is indicated by its ISO 639 code.
Visualization of the language embeddings
Outline

1. Motivation
2. Model
3. Learning
4. Experiments
5. Conclusion
Conclusion

- We represent language identities with language embeddings and use them to encode language similarity.
- The language embeddings are used for grammar parameter prediction and auxiliary language identification task.
- The language embeddings learned in our model can capture language similarity that can not be inferred from phylogenetic knowledge.
Multilingual Grammar Induction with Continuous Language Identification

Wenjuan Han, Ge Wang, Yong Jiang, Kewei Tu

ShanghaiTech University, Shanghai, China
Alibaba Group
{hanwj, wangge, tukw}@shanghaitech.edu.cn
{yongjiang.jy}@alibaba-inc.com

November 9, 2019