A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages

Clara Vania¹, Yova Kementchedjhieva², Anders Søgaard², and Adam Lopez¹

c.vania@nyu.edu, yova@di.ku.dk, soegaard@di.ku.dk, alopez@inf.ed.ac.uk

¹ILCC, School of Informatics, University of Edinburgh
²University of Copenhagen, Copenhagen, Denmark

Realistic Low-Resource Dependency Parsing
- Few resources — no taggers (POS or morphological) are available.
- Parsers must learn from words or characters only.

SCENARIOS: What can we do ...

S1: with a very small target treebank for a low-resource language?
S2: if we also have a source treebank for a related high-resource language?
S3: if the source and target treebanks do not share a writing system?

PARSING STRATEGIES

DATA AUGMENTATION [S1, S2, S3]
- Tree Morphing (Morph; Sahin and Steedman, 2018)
  
  ![Original sentence](image)
  ![Cropped sentence](image)
  ![Rotated sentence](image)

  Figure 1: Operations on the sentence “She wrote me a letter”.

- Nonce Sentence Generation (Nonce; Gulordava et al., 2018)

  "He borrowed a book from a library."

  ✓ He bought a book from the shop.

   X He wore an umbrella from the library.

CROSS-LINGUAL TRAINING [S2, S3]
1. Train a multilingual model using the source and target treebanks.
2. Fine-tune by training the model further only on the target treebank.

TRANSLITERATION [S3]
When the source and target treebanks do not share a writing system, we can map them into the same ‘pivot’ alphabet.

EXPERIMENTS & RESULTS

Parsing model: a neural transition-based dependency parser, with soft parameter sharing on words and characters. (de Lhoneux et al., 2018)

Parsing North Sámi

Data augmentation helps generate up to 4-5 times more training data.

Parsing truly low-resource languages

RESULTS

Comparison to CoNLL 2018 shared task (best set)

<table>
<thead>
<tr>
<th>Language</th>
<th>80</th>
<th>64</th>
<th>48</th>
<th>32</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galician</td>
<td>66.2</td>
<td>74.3</td>
<td>70.5</td>
<td>69.2</td>
<td></td>
</tr>
<tr>
<td>Kazakh</td>
<td>24.2</td>
<td>31.9</td>
<td>25.5</td>
<td>29.2</td>
<td></td>
</tr>
</tbody>
</table>

CONCLUSION

- S1: linguistically motivated data augmentation is helpful.
- S2: cross-lingual training gives the best improvement, but data augmentation still helps.
- S3: transliterating treebanks to a common orthography is very effective.