Choosing Transfer Languages for Cross-Lingual Learning
Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shrutí Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell*, Graham Neubig

Language Technologies Institute, Carnegie Mellon University; *National Research Council, Canada

Cross-lingual Transfer

• Better learning for low-resource task languages by training on related high-resource transfer language

• Question 1: What transfer language do we choose?
  ○ e.g., when processing Azerbaijani, do we choose French, with lots of data, or Turkish with similar typology?

• Question 2: Can we do this process automatically?
  ○ Given a new task language, choose a transfer language without training.

Experimental Paradigm

1. Perform large-scale experiments, exhaustively transferring from/to many different languages.

2. Learn a ranking model, LangRank, to recommend transfer languages, from best to worst, for a new task language.

3. From analysis, gain insights into what makes a good transfer language.

LangRank Model for Choosing Transfer Languages

• Two Categories of Features
  ○ Linguistic features (dataset independent)
    ▪ From URIEL Typological Database
    ▪ Distance between languages deriving from geography, language family tree, phonological features, and syntactic structures
  ○ Statistical features (dataset dependent)
    ▪ Dataset size
    ▪ Word overlap: \( \frac{\sum \text{corpora}}{\text{tokens}} \)
    ▪ Type-token ratio (TTR):

• NLP tasks, dataset, # task languages/# transfer languages
  ○ Machine translation (MT), TED talk corpus, 54/54
  ○ Entity linking (EL), Rijhwani 2019 corpus, 9/53
  ○ Part-of-speech tagging (POS), Universal Dependencies v2.2, 26/60
  ○ Dependency parsing (DEP), Universal Dependencies v2.2, 30/30

2. Ranking Evaluation in NDCG@3

<table>
<thead>
<tr>
<th>Method</th>
<th>MT</th>
<th>EL</th>
<th>POS</th>
<th>DEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best baseline by linguistic feature</td>
<td>(d_{\text{gen}})</td>
<td>(d_{\text{gen}})</td>
<td>(d_{\text{gen}})</td>
<td>(d_{\text{gen}})</td>
</tr>
<tr>
<td></td>
<td>24.2</td>
<td>50.9</td>
<td>15.7</td>
<td>46.4</td>
</tr>
<tr>
<td>Best baseline by statistical feature</td>
<td>(o_{\text{SU}})</td>
<td>(o_{\text{W}})</td>
<td>(o_{\text{W}})</td>
<td>(o_{\text{W}})</td>
</tr>
<tr>
<td></td>
<td>29.2</td>
<td>30.7</td>
<td>13.4</td>
<td>52.3</td>
</tr>
<tr>
<td>LANGRANK</td>
<td>53.7</td>
<td>63.0</td>
<td>28.9</td>
<td>65.0</td>
</tr>
</tbody>
</table>

3. Feature Importance for Different NLP Tasks

• Most influential features
  ○ MT: dataset size, word overlap
  ○ EL: geographic, syntactic
  ○ POS: dataset size, TTR
  ○ DEP: geographic, word overlap

4. Closer Look at Examples

• By considering all linguistic and statistical features jointly, LangRank outperforms strong baseline in task aze

• Even in hard case ben where baselines fail, LangRank remains good

• In small EL dataset, LangRank overfits and fails task tel, while linguistic features serve as good fallback

5. Best Performance Attainable in Top K Recommendations

• If we train models on the top K transfer languages suggested by the ranking model, and pick the best one, how good a model do we expect to get?