**INTRODUCTION**

- A standard strategy for comparing competing theories of learning in morpho-phonology is to compare the fit of models that embody each theory to data.
- This strategy can be used to study the role of bias in learning, comparing models with UG bias to domain-general models.
- It is important to use the strongest available domain-general models as the baseline for comparison.
- Furthermore, as better domain-general learning mechanisms are discovered, old model comparisons should be re-examined.

**Case Study: Romanian Plurals**

- Romanian has four plural suffixes: (-i -uri -a1e). Predicting which suffix pluralizes a given neutral or feminine stem is difficult.
- Previous best performance achieved by hand-tuned Conditional Maximum Entropy Grammar by Grossu & Wilson (2016): $p$(suffix|stem) $\propto \exp\left(- \sum_{c \in C} w$(suffix, stem)$)$
- Network directly predicts plural form given singular input: (f: fata $\rightarrow$ fete ), (m: scaun $\rightarrow$ scaune )

Given random 80/20 train/test split of the 39,500 singular/plural pairs collected by Grossu & Wilson:

<table>
<thead>
<tr>
<th>Romanian</th>
<th>Network</th>
<th>MaxEnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feminine</td>
<td>95.8%</td>
<td>92.2%</td>
</tr>
<tr>
<td>Neuters</td>
<td>86.2%</td>
<td>80.3%</td>
</tr>
</tbody>
</table>

**Case Study: English Past**

- Humans can generalize sub-regularities in irregular past tense forms (swim/swam/swum – spring/sprung/sprung).
- Albright & Hayes (2002) trained their Minimal Generalization Learner to predict past forms of 4253 CELEX stems.
- They used the model to predict human language production probabilities of past tense forms given ‘wug’ present stems.
- Network trained on the same CELEX data correlates with human production probabilities better than MGL.

<table>
<thead>
<tr>
<th>English</th>
<th>Network</th>
<th>MGL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular (rie $\rightarrow$ ried, n=58)</td>
<td>.735</td>
<td>.619</td>
</tr>
<tr>
<td>Irregular (rie $\rightarrow$ rofe, n=74)</td>
<td>.711</td>
<td>.143</td>
</tr>
</tbody>
</table>

**Conclusions and Outstanding Issues**

Modern recurrent neural networks provide a strong domain-general morpho-phonological learning baseline. This makes them a promising tool for studying the role cognitive and linguistic biases play in learning:

- Are there attested patterns that networks have a hard time learning (e.g., as shown by slower convergence during training)?
  - Of particular interest here would be reduplication and metathesis, behaviors that can’t be easily represented by simple finite-state transducers without extensive restrictions.
  - If such patterns exist, it is possible to bias the network to make learning easier (e.g., by initializing weights at a specific starting point).

- Similarly, are there unattested or difficult patterns, as found either by typological surveys or artificial grammar learning experiments, that networks learn too easily?
  - If so, what kind of bias or regularization can be built into the networks to limit learnability?

- What representations and mechanisms do networks actually learn?
  - Do hidden units show large changes in activation when particular informative input symbols are reached (Kirov et al. 2011; Kadar et al. 2016)? If so, network analysis can help us understand and describe phonological patterns by highlighting important parts of the input that we had not considered to be relevant cues.
  - Can learned representations be transformed into feature vectors comprised of traditional linguistic features (e.g., +/- vocalic, +/- labial)?

**References**