You don’t need language-specific tuning.

**Multilingual**, sparse models are very competitive.

---

**DeepSPIN @ SIGMORPHON**

Ben Peters† and André F.T. Martins‡
†Instituto de Telecomunicações
‡Unbabel

---

**One size (and model) fits all**
- Per-language hyperparameter tuning is expensive.
- Small train sets require extra (often artificial) data.
- Multilingual training solves both problems.

**Sparse seq2seq**
- Just replace softmax with entmax everywhere.
- Interpretable sparse attention.
- Sparse output distributions can make decoding exact.
- Requires no other changes to architecture.

**Multilinguality**
- Label each sample with its language.
- Learn a language embedding for each label.
- Concatenate to the character embedding at each step.

---

**Phonology-aware char embeddings**
- Multilingual g2p maps from disjoint scripts to shared IPA.
- Grapheme embeddings cluster by phonological similarity.
- Applications to transliteration?

---

**How data-hungry is inflection?**

---

**Inflection results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc.</th>
<th>Lev. Dist.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inflection-entmax-1.5</td>
<td>90.5</td>
<td>0.217</td>
</tr>
<tr>
<td>Inflection-sparsemax</td>
<td>90.9</td>
<td>0.211</td>
</tr>
<tr>
<td>Baseline</td>
<td>90.6</td>
<td>0.215</td>
</tr>
</tbody>
</table>

- Tied for first place!
- All models use multi-encoder RNNs
- Same model as DeepSPIN’s 2019 submission

**g2p results**

<table>
<thead>
<tr>
<th>Model</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN-entmax-1.5</td>
<td>14.47</td>
<td>2.85</td>
</tr>
<tr>
<td>RNN-sparsemax</td>
<td>14.19</td>
<td>2.78</td>
</tr>
<tr>
<td>Transformer-entmax-1.5</td>
<td>14.15</td>
<td>2.92</td>
</tr>
<tr>
<td>Transformer-sparsemax</td>
<td>14.53</td>
<td>2.92</td>
</tr>
<tr>
<td>Baseline (RNN)</td>
<td>16.84</td>
<td>3.99</td>
</tr>
</tbody>
</table>

- Third place
- Best-performing transformer

---

**What is entmax?**

\[
\alpha\text{-entmax}(z) := \text{argmax}_{p \in \Delta^d} \frac{z^\top p + \frac{1}{\alpha} \sum_j (\frac{p_j}{\alpha} - z_j)^2}{\alpha}, \quad \alpha \neq 1
\]

where

\[
H^{\alpha}(p) = \begin{cases} \frac{1}{\alpha(\alpha-1)} (p_1^2 - p_1^2), & \alpha \neq 1, \\ \frac{1}{\alpha} \sum_j p_j^2, & \alpha = 1 \end{cases}
\]

- \(\alpha = 1 \rightarrow\) softmax
- \(\alpha > 1 \rightarrow\) sparsity possible
- \(\alpha = 2 \rightarrow\) sparsemax
- \(\alpha = \infty \rightarrow\) argmax
- Sparse, differentiable softmax replacement.