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

martins + <en> → mɑɹtɪnz

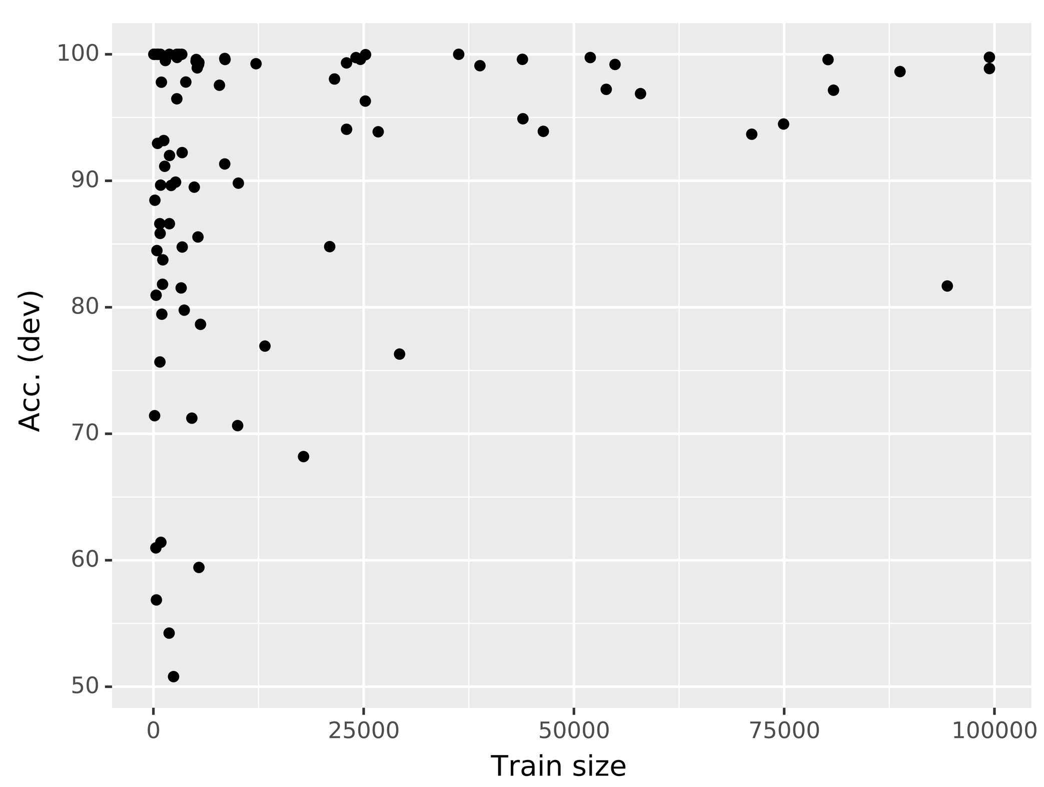
martins + <pt> → mɐrtĩʃ

- 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-sparsemax dev results; each point is a language.
- Bigger is better, but other factors make a huge difference.
- **Typology** matters, but it doesn't explain the sometimes large differences between related languages.

Inflection results

Model	Acc. ↑	Lev. Dist. ↓
Inflection-entmax-1.5	90.5	0.217
Inflection-sparsemax	90.9	0.211
Baseline	90.6	0.215

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

g2p results

Model	WER ↓	PER ↓
RNN-entmax-1.5	14.47	2.85
RNN-sparsemax	14.19	2.78
Transformer-entmax-1.5	14.15	2.92
Transformer-sparsemax	14.53	2.92
Baseline (RNN)	16.84	3.99

- Third place
- Best-performing transformer

What is entmax?

$$\alpha\text{-entmax}(\boldsymbol{z}) := \underset{\boldsymbol{p}\in\triangle^d}{\operatorname{argmax}}\, \boldsymbol{p}^\top \boldsymbol{z} + \mathsf{H}^{\mathrm{T}}_\alpha(\boldsymbol{p}) = [(\alpha-1)\boldsymbol{z} - \tau\mathbf{1}]_+^{1/\alpha-1}$$

where

$$\mathsf{H}^{\mathrm{T}}_\alpha(\boldsymbol{p}) := \begin{cases} \frac{1}{\alpha(\alpha-1)}\sum_j \left(p_j - p_j^\alpha\right), & \alpha \neq 1, \\ \mathsf{H}^{\mathrm{S}}(\boldsymbol{p}), & \alpha = 1 \end{cases}$$

.

- $\alpha = 1 \to$ softmax
- $\alpha > 1 \to$ sparsity possible
- $\alpha = 2 \to$ sparsemax
- $\alpha = \infty \to$ argmax
- Sparse, **differentiable** softmax replacement.

← paper

instituto de telecomunicações

DeepSPIN ("Deep Structured Prediction in NLP") is a research project funded by the European Research Council.

erc