Phonotactic learning with neural language models

Connor Mayer
Max Nelson
connormayer@ucla.edu
manelson@umass.edu
SCiL 2020

Introduction

Humans display gradient preferences towards unattested sequences of sounds.

- Phonotactic models that predict gradient preferences give insight into computations and representations (Hayes and Wilson, 2008; Albright, 2009; Daland et al, 2011; Futrell et al, 2017).
- Existing models operate on some form of *n*-grams.

This task is similar to the more general task of LANGUAGE MODELING.

- Language models assign probabilities to sequences.

Recurrent neural language models outperform *n*-gram variants (Elman, 1990; Mikolov et al., 2010; Sundermeyer et al. 2012).

- These models overcome some limitations of *n*-gram models.

Goal: Show that RNN architectures can be adapted to serve as phonotactic models, providing a closer match to human judgements than existing models. Use these models to probe questions of representation and claims of poverty of the stimulus.

Model architecture

Simple Recurrent Neural Networks (Elman, 1990) – define a probability distribution over phonemes conditioned on the preceding sequence of phonemes.

- Network's state depends only on the current input and the previous state:

$$h_t = tanh(W_x x_t + W_h h_{t-1} + b_h)$$

- Probability distribution over the next phoneme:

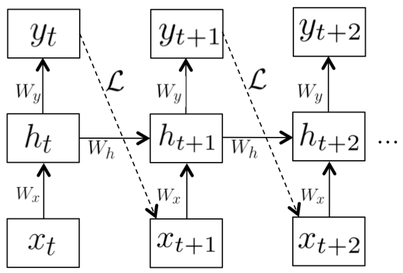
$$\hat{y}_t = (W_y h_t)$$

- Model is trained to optimize predictions to match observed sequences:

$$L(y, \hat{y}) = -y \cdot log(\hat{y})$$

Two types of phoneme representations (x_t):

- FEATURES: Ternary phonological feature vectors.
- EMBEDDINGS: Randomly initialized vectors in $\mathbb{R}^n$, learned along with weights.



Evaluation data sets

Three data sets with different phonotactic properties:

1. Finnish vowel harmony – Unbounded dependencies.
2. English sonority projection – Poverty of the stimulus generalization.
3. Cochabamba Quechua – Restriction involving an unnatural class.

Hayes & Wilson phonotactic learner as a baseline (H&W; Hayes & Wilson, 2008).

H&W scores novel forms with MAXENT VALUE:

$$P_{HW}(x) = \exp\left(-\sum_{i=1}^{N} w_i C_i(x)\right)$$

RNNLM scores novel forms with PERPLEXITY:

$$(x) = \exp\left(-\sum_{i=1}^{|x|} \frac{1}{|x|} log_2(p(x_i))\right)$$

Finnish vowel harmony

Words in Finnish generally contain all front {y, ø, æ} or all back {u, o, a} vowels.

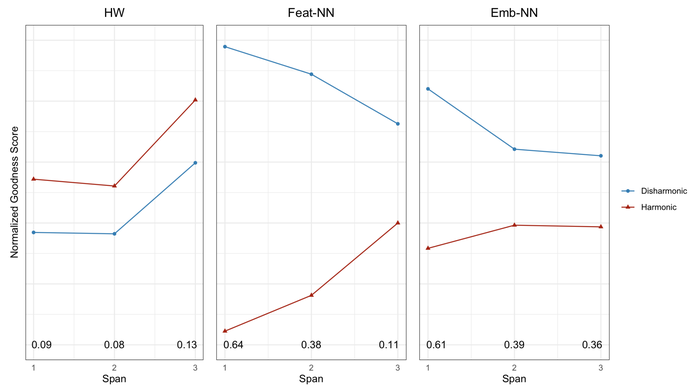
- {i, e} are TRANSPARENT to harmony.

Are RNNs more effective than H&W in learning this long-distance dependency?

Trained on corpus of 94k Finnish words.

- Tested on 20k nonce words: 10k harmonic and 10k disharmonic.

Model scores of nonce words by span of consecutive transparent vowels:



Neural models distinguish between harmonic and disharmonic forms more robustly (Cohen's *d*), even as transparent span increases.

English sonority sequencing

Cross-linguistic preference for syllables that conform to SONORITY SEQUENCING PRINCIPLE (SSP): monotonically rising sonority in onset, falling in coda.

- Innate (Berent et al., 2007; 2008) or learned from data (Daland et al., 2011)?

English sonority sequencing (cont.)

Trained on 133k word CMU dict.; tested on nonce words (Daland et al., 2011).

- Attested, unattested, and marginal onset clusters of varying sonority profiles.

Model score correlations with human judgements from Daland et al. (2011), by cluster type.

	Overall	Attested	Unattested	Marginal
H&W (H)	0.759	0.000	0.686	0.362
Feat	0.868	0.354	0.823	0.551
Emb	0.853	0.491	0.738	0.664

RNN models learn English SSP without phonetic features or syllable structure.

- Featural model generalizes better: embedding model is overfitting?

Cochabamba Quechua

CQ exhibits laryngeal co-occurrence restrictions: ejective and aspirated stops must be the FIRST stop in the root (Gallagher, 2019).

- Plain stops can occur after any type of stop.
- /q/ is always realized as [ĸ], but still patterns with the plain stops.

Do models with learned embeddings do better with this pattern?

Trained on corpus of 2,500 root forms.

Tested on nonce forms from Gallagher (2019):

- 25 licit: e.g., [sap'a].
- 25 illicit by [k]: e.g., [*kap'a].
- 25 illicit by [ĸ]: e.g., [*ĸap'a].

	Licit	Illicit (k)	Illicit (ĸ)
H&W (P_{HW})	0.67	0.28	0.30
Feat ( )	4.91	8.45	7.42
Emb ( )	4.89	8.45	7.55
Tied Emb ( )	4.91	8.28	7.16

No models make a significant distinction between k-initial and ĸ-initial forms.

- Because of the unnatural class, H&W must represent the restriction with multiple independent constraints (conspiracy).
- The embedding model has the ability to learn an unnatural class. Does it?

Cosine similarity between [ĸ] and mean representation of continuants and non-continuants.

	continuant	non-continuant
Featural [ĸ]	0.62	0.51
Emb [ĸ]	−0.26	0.19

Embedding model treats [ĸ] more like a non-continuant.

Discussion

RNN phonotactic models perform at least as well in matching human judgements as existing phonotactics. They have several desirable properties:

- Overcome limits of *n*-gram models (Finnish).
- Provide insight into what information is present in learning data (English).
- Provide insight into possible distinctions between representations and processes that operate on them (CQ).

Future questions:

- Do humans learn unnatural classes like H&W or like the embedding models?
- How can we gain insight into what RNNs have learned?