Learning and generalizing phonotactics with recurrent neural networks

Max Nelson
Connor Mayer
manelson@umass.edu
AMP 2019, Stony Brook University

Introduction

Humans display gradient preferences towards unattested sequences of sounds

- Phonotactic models that predict gradient preferences can give insight into computations and representations (Hayes and Wilson, 2008; Albright, 2009; Daland et al, 2011; Futrell et al, 2017)
- One such preference is sonority sequencing
  - ⇒ Crosslinguistically attested preference for onset clusters which increase in sonority
  - ⇒ Is a built in bias towards certain sonority profiles necessary to account for observed sonority sequencing effects?

Goal: Can gradient human sonority sequencing preferences be learned from lexical statistics alone?

Background

Not the first with this question (Berent et al 2007, 2008; Albright, 2007; Ren et al, 2010; Daland et al 2011; Jarosz and Rysling 2017)

- Daland et al. (2011) collect human judgements, train phonotactic models on CELEX, check correlations between model and human judgements
  - ⇒ Run on syllabified and unsyllabified data
  - ⇒ Best result: HW phonotactic learner (Hayes and Wilson, 2008)
- Correlations with aggregate human judgements of words containing attested, unattested, and marginally attested onsets of varying sonority profiles

Onsets			Tails
Attested	Marginal	Unattested	
tw tr sw	gw ʃl	pw zr mr	-ɑtɪf
ʃr pr pl	vw ʃw	tl dn km	-ibɪd
kw kr kl	ʃn ʃm	fn ml nl	-ɑsɪp
gr gl fr	vl bw	dg pk lm	-ɛpɪd
fl dr br	dw fw	ln rl lt	-igɪf
bl sn sm	vr θw	rn rd rg	-ɛzɪg

- Best results from key models:

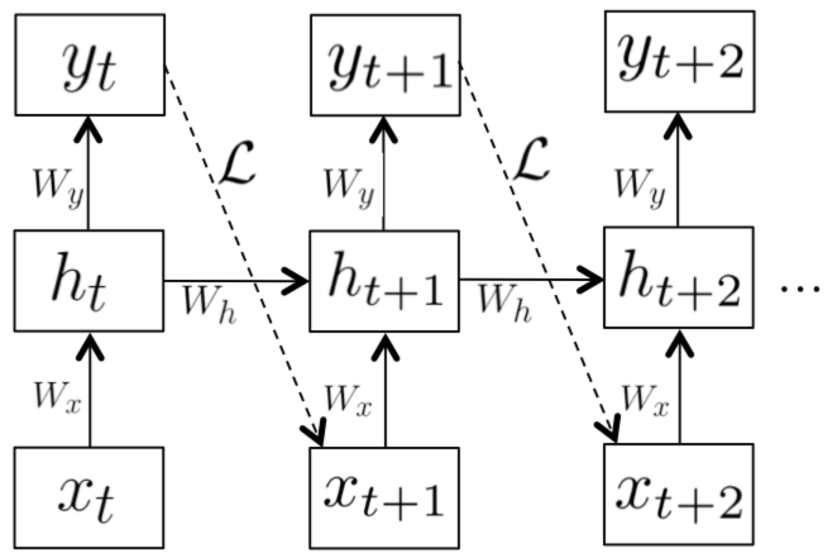
Model	Overall	Attested	Marginal	Unattested
BH	0.24	0.30	0.22	-0.26
HW	0.80	0.00	0.00	0.70
HW[syll]	0.83	0.00	0.02	0.76

- To generalize models must represent similarity between segments
- All models perform better on syllabified data
- Projection is learnable from lexical statistics provided featural representations and syllabification
- Aside: This result has been shown to not hold for Polish (Jarosz, 2017)

Secondary goal: Can sonority sequencing preferences be learned with unsyllabified data and without prespecified linguistic features?

Neural Language Models

- Language modeling - defining a probability distribution over sequences, operationalized as next element prediction
- Elman (1990) - sRNNs to predict upcoming segment, allow probability to be conditioned on entire preceding sequence
- Bengio (2003) - Continuous representations in neural language models
  - ⇒ Random real valued representation, optimized with objective, distributional information
- Mikolov (2010) - Continuous representations in RNN language models
- Mirea and Bicknell (2019) - Continuous representations for next phoneme prediction with LSTMs



Current Approach

- HW and several sRNN LMs trained on 133,000 word CMU dictionary, no syllable annotation
- Fit models used to make predictions for all items in Daland et al.'s experiment, evaluated by measuring linear correlation between model and human judgement
- Two different phoneme representations, *features* and *embeddings*
- **Feature models** - fixed vector 26 ternary features (Hayes, 2009)
- **Embedding models** - randomly initialized vector in $\mathbb{R}^{24}$
- All models trained on next phoneme prediction, optimizing cross-entropy

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

- Input and output embeddings are optionally tied (Press and Wolf, 2018)
- Hyperparameters selected by grid search on 70/30 split of CMU dict.

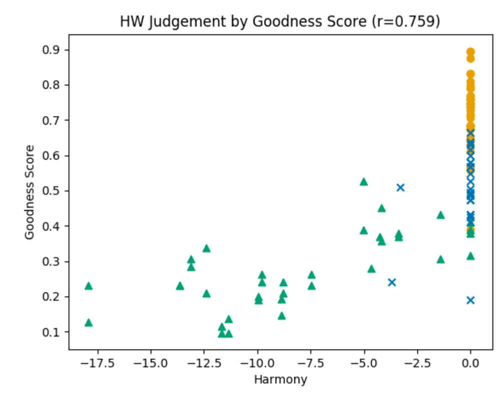
Predictions

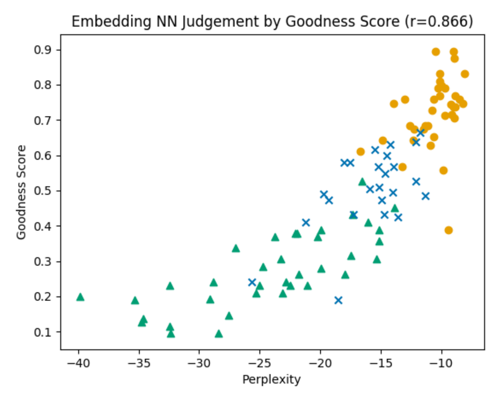
1. Neural models will be able to learn and generalize sonority sequencing as well as existing models
2. Embedding models will learn representations that capture sonority classes and predict sonority projection

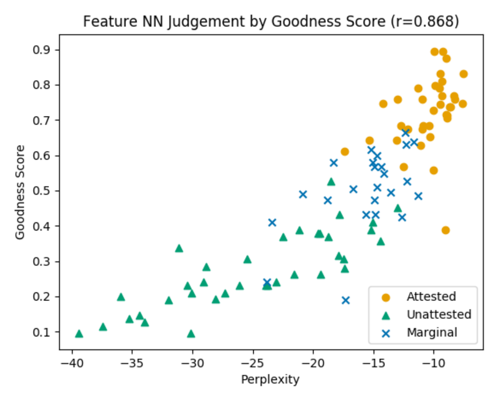
Results

- Correlation coefficients between human and average model judgement

	Overall	Attested	Unattested	Marginal
H&W	0.759	0.000	0.686	0.362
Feat	0.868	0.354	0.823	0.551
Emb	0.866	0.365	0.765	0.609
Tied Emb	0.853	0.491	0.738	0.664

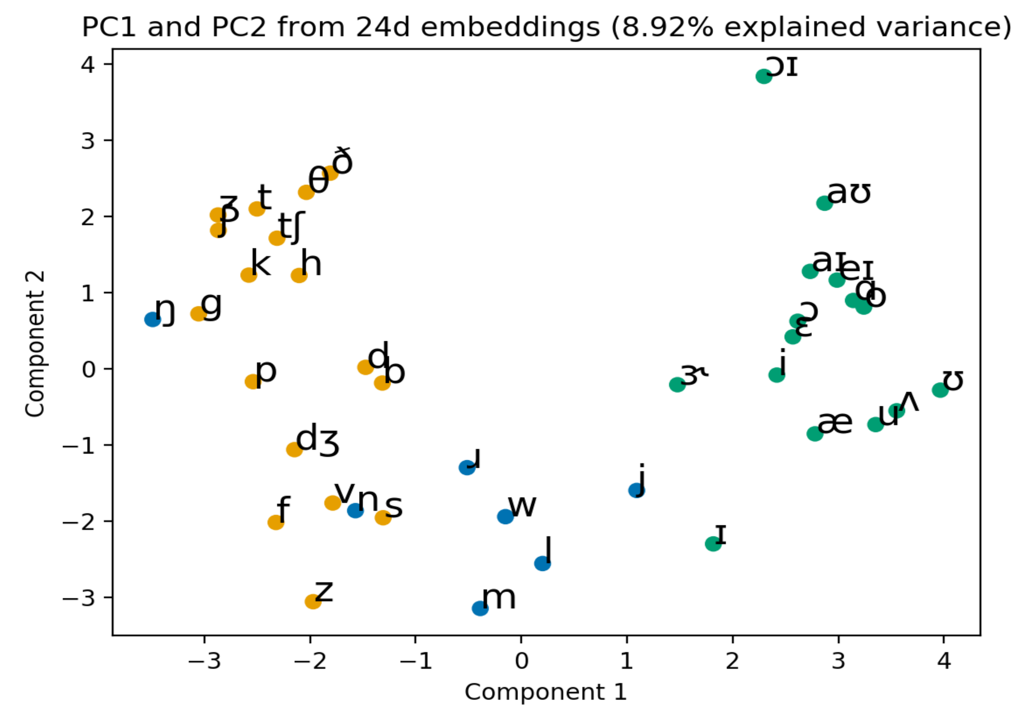






Are the embeddings capturing phonological features?

- PCA of tied embeddings - separation of sonorants, obstruents, and vowels



- **Probe task**: Can a 1-layer softmax classifier identify feature specifications from embeddings?
- 1000 classifiers for any feature with at least 7 positives and negatives

	Avg p(correct)		
Feature	Positive	Negative	Overall
SYLLABIC	0.981	0.970	0.975
CONSONANTAL	0.988	0.914	0.951
SONORANT	0.823	0.927	0.875
VOICE	0.666	0.645	0.655
CONTINUANT	0.469	0.392	0.431
ANTERIOR	0.490	0.702	0.596

Conclusion

- Neural models predict sonority projection, also make gradient predictions for attested onsets
- Distributional features predict behavior pretty well - but linguistically informed features better predict generalization