Learning and generalizing phonotactics with recurrent neural networks

Introduction - A central problem for modeling phonotactic learning comes from the fact that models need to display gradient preferences for unattested sequences if they are to be said to capture human behavior. One domain in which this problem is particularly salient is in sonority projection. There have been a number of studies, spanning multiple unrelated languages, that illustrate a preference towards unattested complex onsets which rise in sonority over those that have a flat or falling sonority profile[1].

The reliability of sonority sequencing effects on phonotactic judgements has lead some to hypothesize that there must be a universal, potentially innate, preference towards certain sonority profiles (Berent et al., 2008). Others have proposed that an innate bias is not needed to capture human behavior. Daland et al. (2011) reject the innatist hypothesis by showing that computational models are able to predict English speaker sonority projection effects based only on the lexical statistics of English when two conditions are met: (1) the model must have a rich featural system that captures sonority levels and (2) it must be trained on data that is annotated for syllable structure.

The current work presents a neural network phonotactic learning model which supports the idea that English sonority projection can be learned by lexical statistics, while relaxing Daland et al.'s condition (1) and removing condition (2). Neural models are shown to better capture sonority projection than existing models despite not being provided with syllable structure annotations and, in one case, not being provided with phonological features.

Recurrent neural network phonotactic learning - The present model is based on recurrent neural networks (RNNs), which were designed to handle the problem of sequentially ordered inputs of varying length. RNNs process input sequences one element at a time and, at every timepoint, use the current element in the input to compute a hidden state, and then use that hidden state to compute an output for the current timepoint. The hidden state is then combined with the input of the next timepoint (Elman, 1990). At any timepoint, the hidden state can be thought of as the network's representation of the preceding sequence.

In the approach taken here, a simple RNN is trained to predict the upcoming phoneme given all preceding phonemes in the same word. Similar to n-gram models, the network is predicting a probability distribution over phonemes given its knowledge about the preceding sequence. Unlike n-grams, the network is able to represent long-distance dependencies by learning what information to pass forward through the hidden state. The models are trained on the CMU dictionary to simulate a lexicon; data is not annotated for syllable structure.

Two versions of an RNN phonotactic learner are presented. In the first, phonemes are represented with a set of 23 standard ternary (positive, negative, or unspecified) phonological features taken from Hayes and White (2013). In the second, phonemes are represented with randomly initialized, real-valued, 24-dimensional vectors that are adjusted during training by backpropogating errors to the input. This reflects the standard NLP approach to representing words with embeddings (Bengio et al., 2003). Following Press and Wolf (2018), input and output representations are tied. This approach relaxes the stipulation made by Daland et al.: the learner is not provided with featural representations that capture sonority tiers, but it it does have the ability to learn sonority classes if there is evidence for them in the lexicon.

Daland et al's experimental results - Along with testing a number of computational models, Daland et al. report the results of an experiment in which English speakers were

[1]Korean: (Berent et al., 2008); Mandarin: (Ren et al., 2010); English: (Daland et al., 2011); Polish: (Jarosz and Rysling, 2017)

asked to judge the acceptability of novel words which were formed by pairing attested, marginally attested, and unattested clusters of varying sonority profiles with novel suffixes. Daland et al. evaluate existing computational models by computing the trained model's judgement of these words and then testing the extent to which model judgements correlate with human judgements. The results of Daland et al.'s experiment are publicly available and are used to evaluate the proposed model.

Results - Correlation coefficients between model judgments and Daland et al.'s experimental results are shown below. Following Daland et al., separate correlations are reported for attested, marginally attested, and unattested clusters alongside the overall correlation coefficient. Models below the single horizontal line, including the Hayes-Wilson phonotactic learner (2008), the generalized neighborhood model (Bailey and Hahn, 2001), the phonotactic probability calculator (Vitevitch and Luce, 2004), and a baseline syllable bigram model, are implemented by Daland et al. and trained on syllable structure annotated data.

Model	Attested	Marginal	Unattested	Overall
Feat. RNN	0.23	0.40	0.68	0.81
Emb. RNN	0.37	0.68	0.76	0.86
Hayes-Wilson	0.00	0.02	0.76	0.83
GNM	0.32	0.23	-0.22	0.31
Vitevitch-Luce	0.30	0.06	0.27	0.56
Syllable Bigram	0.19	0.16	0.22	0.78

The neural model with randomly initialized phoneme embeddings better predicts human behavior than all other models, despite not having syllable structure annotated data and not being provided with phonological features that represent sonority. The featural RNN performs comparably to the best previous model, a version of the Hayes-Wilson learner.

The first step towards determining why the embedding model does so well is probing the content of the learned embeddings. A set of binary single-layer softmax classifiers were trained to predict whether a phoneme had a positive or negative value for an SPE-style phonological feature given its embedding. Because of the small sample size for any given feature, 1000 models were fit for every feature which had more than 6 positive and 6 negative examples with a randomly selected pair held out as a test case. The results indicate that features which are important in determining sonority ([SYLLABIC], [CONSONANTAL], [SONORANT]) are well encoded in the embeddings while those which less directly represent sonority ([VOICE], [CONTINUANT], [ANTERIOR]) are not encoded in the embeddings.

Conclusions - The neural models presented here are able to model sonority projection as well or better than existing phonotactic learning models without syllable structure annotated data and, in the case of the embedding model, without being provided with a set of phonological features which explicitly encode the sonority hierarchy. This result strengthens Daland et al.'s claim that English sonority projection can be learned from lexical statistics alone by showing that not only can sonority projection be learned by a statistical learner with no innate bias, but also that the stipulations made by Daland et al. regarding the success of statistical learners can be relaxed or removed. Neural network phonotactic models are able to learn sonority projection without syllable structure annotated data and without a prespecified feature set.