Learning Phonotactics in a Differentiable Framework of Subregular Languages

Huteng Dai (Rutgers), Richard Futrell (UC Irvine)

Most phonotactics patterns are argued to fall in subregular regions in Chomsky Hierarchy such as Strictly Piecewise (SP) and Strictly Local (SL) languages (Rogers & Pullum 2011). There has been a growing interest in linking the study of subregular languages and probabilistic models, which not only fills the gap in Formal Language Theory, but also sheds light on the structural nature of phonological acquisition (Dai 2021; Shibata & Heinz 2019). Here we take up the challenge of establishing a differentiable probabilistic framework to facilitate the comparison of subregular languages and the application of machine learning methods such as *backpropagation* to learn phonotactic constraints from noisy corpus data. Inspired by works on Probabilistic Finite-state Automata (Vidal et al 2005; PFA), we present a linear-algebraic formulation of PFA which enables learning the structure of various subregular languages, including SL, SP, and SP + SL. We evaluate our learner on corpus data from Navajo and Quechua, with an emphasis on the ability to learn nonlocal constraints, such as *s...ʃ in Navajo.

1. PFA. A PFA consists of a finite set of states Q, an inventory of symbols Σ, an **emission distribution** which gives the probability of generating a symbol $x \in \Sigma$ given state $q \in Q$, and a **transition distribution** which gives the probability of transitioning into new state q' from state q after emission of symbol x. The generation of a particular word $\mathbf{x} = x_1 x_2 \ldots x_n$ works by starting in a particular distinguished **initial state** q_0, generating a symbol x, transitioning into the next state q', and so on recursively until reaching a **final state** q_n.

We parameterize a PFA using a family of matrices. The **emission matrix** (**E**), of shape $|Q| \times |\Sigma|$, gives the probability of emitting a symbol x given a state. Each row in the matrix represents a state, and each column represents an output symbol. Each symbol x is associated with a matrix in **transition matrices** (**T**), of shape $|Q| \times |Q|$. Given a parameterized PFA, the likelihood of a sequence, marginalizing over all trajectories through states, can be efficiently calculated by a forward algorithm following Vidal et al. (2005).

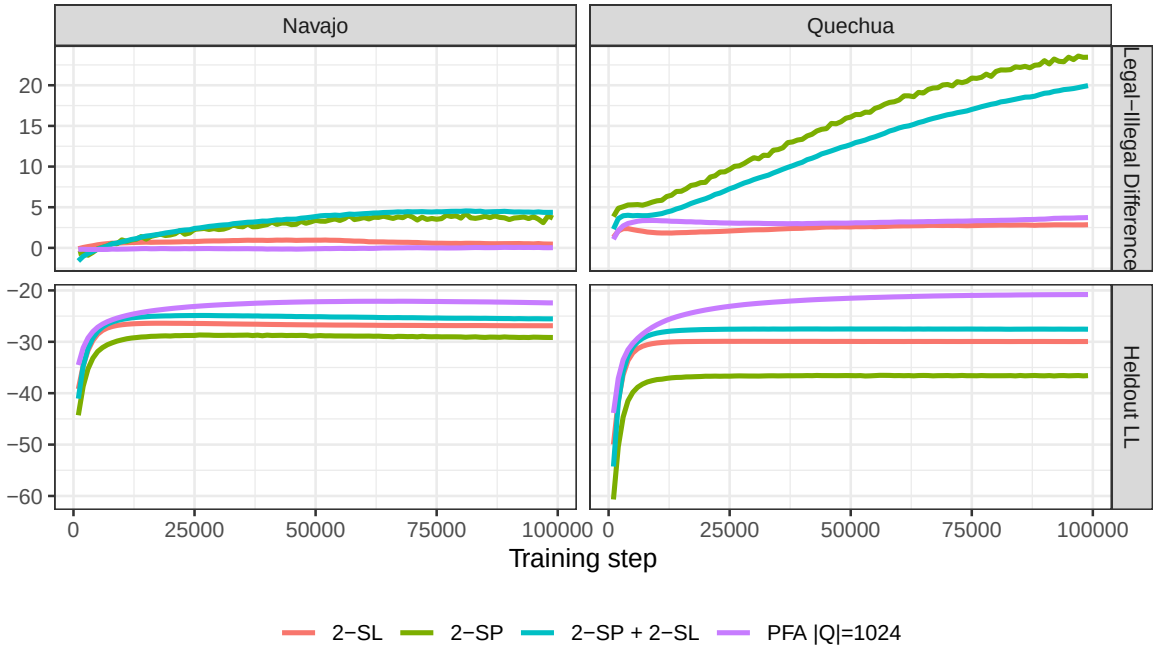
In this construction, the likelihood of a sequence is a differentiable function of the matrices **E** and **T** that define the PFA. This ensures that a learner can compute the gradient of the likelihood of a form with respect to the matrices **E** and **T** and induce a PFA by using gradient descent to search for matrices which maximize the likelihood of training forms, or which maximize related objective functions. This method can be used to induce both the **E** and **T** matrices from data, specifying only the maximum number of states $|Q|$ available to the PFA, without specifying the structure of the transition matrices in advance.

2. SL & SP. In a Strictly k-Local (k-SL) language, each symbol is conditioned only on immediately preceding $k-1$ symbol(s); in a Strictly k-Piecewise (k-SP) language, each symbol depends on the presence of any preceding $k-1$ symbols at arbitrary distance. SP can capture the nonlocal dependencies missed in a SL grammar. For a word [sipoʃ], in a 2-SL language, [ʃ] is **only** conditional on its immediately preceding one symbol [o]; in a 2-SP language, however, [ʃ] is conditional on any preceding one symbol, including [s].

We implement automata that generate strings from SL and SP languages by specifying their corresponding transition matrices **T**. For these automata, the learner only optimizes over their emission matrices **E**. It is also possible to train an automaton with the ability to condition on both SP and SL factors by taking the product of SP and SL automata (Heinz & Rogers 2013). We refer to the language generated by such an automaton as SP + SL.

3. Data. The proposed learner is applied to the datasets of Navajo and Quechua (Gouskova & Gallagher, 2020), in which nonlocal phonotactics are attested. In Navajo, the co-occurrence of alveolar and palatal strident is illegal. The learning data of Navajo includes 6279 Navajo phonological words; we divide this data into a training set of 5023 forms and a held-out set of 1,256 forms. The nonce testing data of Navajo consists of 5000 generated nonce words, which were labelled as illegal ($N = 3271$) and legal ($N = 1729$) based on whether the nonlocal phonotactics are satisfied. In Quechua, any stop cannot be followed by an ejective or aspirated stop at any distance. The learning data of Quechua includes 10804 phonological words, which we separate into 8643 training forms and 2160 held-out forms. The testing data of Quechua consists of 24352 nonce forms which were manually classified as legal ($N = 18502$) and illegal ($N = 5810$, including stop-aspirate and stop-ejective pairs).

4. Result. The following plot shows the performance over the course of maximum-likelihood training of a 2-SP automaton, a 2-SL automaton, a 2-SP + 2-SL product automaton, and an unrestricted PFA with 1,024 states. 'Heldout LL' is the average log likelihood (LL) of forms from the held-out set; a higher LL indicates that the model can generalize from its training data, assigning higher probability to attested but unseen forms. 'Legal–illegal difference' is the difference in log likelihood between 'legal' and 'illegal' forms in the nonce test set; a higher value indicates higher sensitivity to the nonlocal constraints. We find that the unrestricted PFA learner achieves the highest heldout LL, with the SP+SL learner outperforming the SL learner. Only the SP and SP + SL learners achieve substantial legal–illegal differences.



5. Conclusion. By implementing a differentiable framework and comparing the learning of (sub)regular languages against corpus data, we show that inducing an unrestricted PFA produces the best fit to naturalistic held-out forms, while the SP model is superior in capturing nonlocal constraints as evidenced in artificial data. The SP + SL model shows strength both in learning nonlocal constraints and in predicting naturalistic held-out forms, supporting the claim in Heinz & Rogers (2013) that SP + SL is advantageous for characterizing natural language phonotactics.