

Learning Phonotactic Constraints from Surface Forms

The foundational work in learning Optimality Theoretic (OT) grammars (Tesar and Smolensky, 1998) took the basic problem to be determining the constraint ranking and armed the learner with a set of constraints, underlying forms and candidate sets. Related work has shown that learning is possible even when the learning data is quite small relative to the possible number of grammars (Lin 2002, Riggle in progress). Despite these welcome results, there has been a growing interest in algorithms that can learn OT grammars without underlying forms (?) and without constraints.

This paper presents some results of a phonotactic learning algorithm that induces phonological constraints from surface forms alone. Hayes (2004) motivates the need for what he terms a ‘pure phontactic learner’ by hypothesizing that a complete phonological grammar is more easily learned in stages: the phonotactics of the target language is learned prior to the alternations. This strategy follows from the observation that children appear to acquire alternations later than the phonotactics, much of which appears to be in place by as early as one year. The learning mechanism described in (Albright and Hayes, 2002), for example, succeeds in part because it has been given phonotactic constraints that cued rule discovery.

In earlier constraint-learning approaches (Ellison, 2001), the learner seeks to match a very concrete constraint template to the observed forms, keeping successful templates. The learning mechanism here, which is based on ideas in Angluin (1982), is given a feature system and a sample of surface forms (as strings of feature bundles) from a target language, and builds a prefix tree (Figure 1) which is then “collapsed” (Figure 2a) This results in a regular expression which is consistent with the observed forms and ‘minimal’ in a particular sense. The choice of collapsing method is essentially a bias that controls the degree and kind of generalizations made. For example, if set too freely, the algorithm generalizes too much and accepts any future word as grammatical, no matter how ill-formed. The procedure can be iterated over different sets of natural classes. Upon each iteration, the resulting regular expression is converted into a set of negative constraints (e.g. Figure 2b).

Under one particular collapsing method, the learner can successfully learn both local and (unlike bigram and trigram models) long-distance phonotactic constraints. For example, when presented with forms like those in (1), the algorithm learns the constraints *CC, *VV, and *Final-C; three constraints sufficient to describe the language which only allows (C)V syllables. When presented with surface forms like those in (2), it successfully learns a constraint prohibiting [-high, +ATR] vowels to be followed anywhere in the word by [-ATR] vowels.

Despite these successes, there are outstanding problems. In its current form, the learner only learns surface true constraints. Additionally, there is a search problem in finding the sets of natural classes to be submitted to the algorithm.

Finally, it is suggested how additional biases might be incorporated into the current learning device. There is much evidence from cross-linguistic similarities, phonology-phonetics interface studies, and from child language studies that there are universal biases in the acquisition of phonological grammars. The ultimate goal of this research is to create a phonotactic constraint learner that embodies these universal biases. This would obviate the need for a universal constraint set which until now has been the only viable alternative.

- (1) a, ka, eta, piko, obuti, tamano

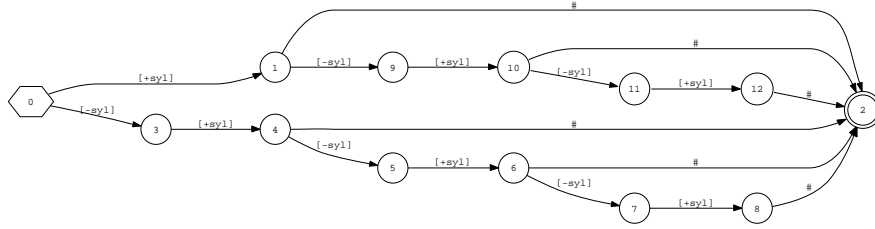


Figure 1: The prefix tree of the forms in (1) with the natural classes [+syl] and [-syl]

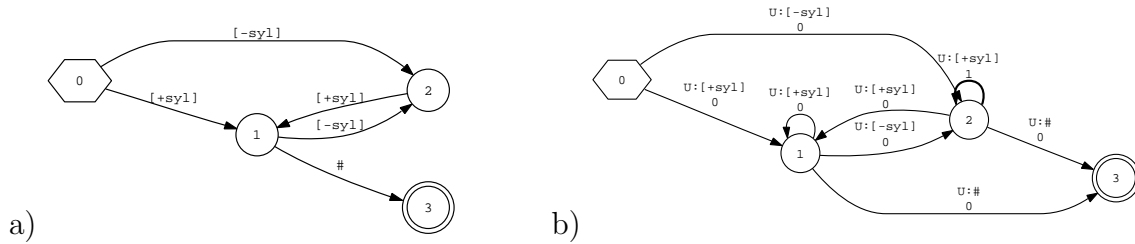


Figure 2: At left is the collapsed regular expression of Figure 1 equivalent to the ((C)V)⁺ language. At right is the *VV constraint extracted from the machine at left.

- (2) ipika epiki opiki Epika Opiki apiki ([a,E,O] are -ATR and
 ipuka epuki opuki Epuka Opuki apuki [i,u,e,o] are -ATR)
 ipeki epeko opeko Epeki Opeki apeko
 ipoki epoko opoko Epoki Opoki apoko
 ipEki EpEka OpEka apEki
 ipOki EpOka OpOka apOki
 ipaki Epaka Opaka apaka

References

- Albright, Adam and Hayes, Bruce. 2002. Modeling English Past Tense Intuitions with Minimal Generalization. Philadelphia: Association for Computational Linguistics. Appeared in Mike Maxwell, ed., Proceedings of the 2002 Workshop on Morphological Learning, Association of Computational Linguistics.
- Angluin, Dana. 1982. Inference of reversible languages. *Journal of the Association for Computing Machinery* 29:741–765.
- Ellison, Mark. 2001. The iterative learning of phonological constraints Association for Computational Linguistics.
- Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the early stages. In *Fixing Priorities: Constraints in Phonological Acquisition*, edited by Rene Kager, Joe Pater, and Wim Zonneveld. Cambridge University Press.
- Tesar, Bruce and Smolensky, Paul. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.