Computational Phonology, class 8: Learning constraints

Adam Albright

CreteLing 2022 — July 2022

SCAN ME

creteling2022.computational.phonology.party

Learning constraints

Three possibilities

- CON is fixed and universal (Prince and Smolensky 1993/2004)
  - Learning problem is reduced to learning URs and rankings
  - Factorial typology
- CON is constructed from universal primitives
  - *Features, Ident(feature), etc.
  - Learner must find constraints, large search space (possibly intractable? Idsardi, Heinz)
  - Factorial typology: expanded, but calculable
- CON is induced based on violation profiles, may contain arbitrary constraints (Doyle et al 2014)

Hayes and Wilson (2008)

Hayes and Wilson (2008): structure of constraints

- Negative: penalize combinations of natural classes
  - $*\begin{bmatrix} -\text{cont} \\ +\text{cor} \end{bmatrix}[+\text{lat}]$, $*[+\text{round}][-\text{round}]$, etc.
- Positive: penalize complement sets
  - $\checkmark s[+\text{nas}] \Rightarrow *\begin{bmatrix} \hat{}+\text{strid} \\ +\text{cont} \\ -\text{voi} \end{bmatrix}[+\text{nas}]$
- Number is exponential in length n
  - $\sum_{1}^{n}(|\text{classes}|^{i} + i(|\text{complement classes}| \times |\text{classes}|^{i-1}))$
- Hayes and Wilson suggest an upper bound on n, depending on number of classes
  - Segments: n=2(ish) (many segmental classes)
  - Stress: n=4 (few stress features)

Searching and including constraints

Two desiderata

- Accuracy: constraints should have few violations in the training data (exceptions)
- Generality: constraints should eliminate large sets of forms

Accuracy: Observed/Expected (O/E)

- Don’t know what combination of constraints can match observed distribution
- Do know that if model predicts a sequence to be common, but it’s rare, then we’re missing a constraint
    - How many violations occur in the data? (Observed)
    - How many violations occur in the set of strings that are consistent with the current grammar? (Expected)

“*when we are seeking a new constraint to add to the grammar, we generate a “sample”—that is, a set of forms drawn from the probability distribution defined by the current grammar.*”

Accuracy: Observed/Expected (O/E) (*cont.*)

- Greedy search: find constraints with greatest discrepancies (O/E closest to 0)
  - If we expect many CC sequences ([pn], [kr], [st], ...) but we never observe them, *CC has very low O/E (0/many)
  - Favor large denominators: upper confidence limit on O/E
- Implementation: start at 0 and gradually increase

Generality

Two heuristics

- Shorter constraints are more general (smaller n)
- Fewer features $\Rightarrow$ larger classes $\Rightarrow$ more general
- Shorter constraints are categorically favored over simpler but longer constraints

Putting this together: a procedure

To find a constraint to add to the current grammar:

- Sample a large set of strings from the probability distribution assigned by the current grammar
- From length $= 1$ to n
  - From highest to lowest generality natural classes
    - Use classes to build a candidate constraint of current length
    - Calculate O/E for candidate constraint
    - If O/E $<$ threshold, break and add to grammar
- Add to grammar: find new weights

The procedure

Example (10), p. 394

```
1   begin with an empty grammar 𝒢
2   for each accuracy level a (increasing O/E)
3       do
4         select the most general constraint with accuracy a
            and add it to 𝒢
5         optimize constraint weights for new grammar 𝒢
6       while a constraint is selected in 4 (and size of 𝒢 < max)
```

You can test the model

- http://www.linguistics.ucla.edu/people/hayes/Phonotactics/index.htm

- See Hayes and Wilson (2008), section 5

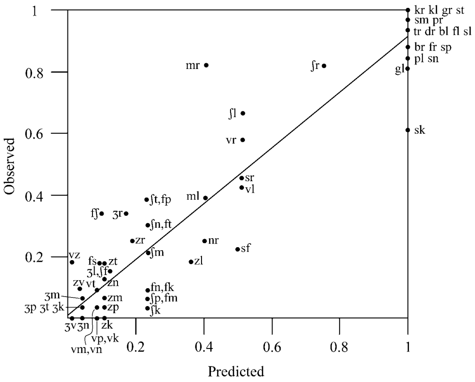

Figure 3
Performance of the model in predicting the data from Scholes 1966

Markedness only models of phonotactics?

- Hayes and Wilson (2008) model: learn markedness constraints only
  - If successful: eliminate (penalize $\Rightarrow$ prob of 0) unobserved structures
- No commitment to repairs

Augmenting the model: variables

- The Hayes and Wilson (2008) model learns constraints over sequences of natural classes
- Many (hand-crafted) phonological constraints in the literature use an additional power: variables
  - Identity: two elements in the string share a feature
  - Distance: an element in a string can take on a range of features
- E.g., Obligatory Contour Principle (OCP) for place: $*C_{[\alpha\text{place}]}\ldots C_{[\alpha\text{place}]}$
  - 'α' encodes identity between two feature values
- Distance
  - '…' encodes any number of intervening segments (= X_0, etc.)

Identity within the string

Berent et al. (2012)

- Hebrew has consonantal roots, generally C-C-C
- OCP restrictions
  - C_1 and C_2 may not be identical
  - C_2 and C_3 may be identical
  - Non-identical consonants with same place generally avoided in both positions
- A blick test (Berent et al 2002)
  - Hebrew speakers judge novel roots with $C_1 = C_2$ worse than $C_2 = C_3$
  - This holds for consonants that exist in Hebrew, and also consonants that don't: /tʃ/, /dʒ/, /θ/

Modeling this generalization

- The Hayes and Wilson learner can mimic OCP constraints with specific feature matrices
  - Place: *[+cor][+seg][+cor], *[+lab][+seg][+lab], etc.
  - Identity: *t[+seg]t (t = [-son,-voi,+cor,…])
  - Or things in between…
- When trained on Hebrew, no reason to posit *θ[+seg]θ, *dʒ[+seg]dʒ
  - High ranking *θ, *dʒ make expected number of θVθ, dʒVdʒ = 0
- Model fails to generalize like humans

Modeling this generalization

- Revised model: allow constraints that index a feature matrix and refer to a copy
  - $*X_i \alpha_i Y$ (repeated X followed by Y)
- Result: models learns $*\#[+seg]\alpha_i$
  - No repeated consonant at beginning of word
- Revised model improves statistical fit to human judgments

Distance: Large window restrictions

- Davis (1984): $*sC_iVC_i$, where C is non-coronal
  - Reflexes across other Germanic languages (Coetzee, 2008)
  - Does not depend on syllable structure: no **spaper* words
- N-gram models with large-n
  - Smoothing, back-off
- Sparsity of the data: generally too few examples of any sequence length $>$4 to clearly differentiate attested from unattested

The learnability problem posed by nonlocal phonology

- When the environment is known to be local, the number of logically possible conditioning environments to explore is proportional to the number of natural classes in the language
- When environments occur at a distance, the number of logically possible environments to be explored rises exponentially with the length of the intervening string

Solving the locality problem: variables

Gouskova and Gallagher (2020): Quechua

Attested combinations				Impossible combinations		
(a) tʃ'uspi	'fly'	(c) rit'i	'snow'	(e) *kup'i	(g) *k'up'i	(i) *kʰup'i
(b) kʰutʃi	'pig'	(d) ʎimpʰu	'clean'	(f) *kupʰi	(h) *k'upʰi	(j) *kʰupʰi

Table 3: Quechua laryngeal restrictions

- "Ejectives and aspirates can only occur non-initially if preceded by fricatives or sonorant consonants"
- Two constraints
    - *[-cont, -son]...[+constricted glottis]
    - *[-cont, -son]...[-cont, +spread glottis]
- Both are non-local, refer to features of stops

Gallagher and Gouskova's approach

- Two things that hold of Quechua
  - Unbounded: *[-cont, -son]…[+constricted glottis]
  - Almost locally bounded: *[-cont, -son]V[+constricted glottis]
- Quechua also bans CCC sequences, so one other thing also holds:
  - *[-cont, -son]C[+constricted glottis]
- Stated even more generally:
  - *[-cont, -son][seg][+constricted glottis]

  In fact, this is what the Hayes and Wilson model discovers for the 'almost locally bounded' constraint, since it is shorter/more general

Gallagher and Gouskova's approach

- Identify cases where the model learns constraints *X[seg]Y
  - These are cases where the identity of the intervening material didn't matter, and perhaps the amount doesn't matter, either
  - E.g., *[-cont, -son][seg][+constricted glottis]
- Posit a new tier: all of the features X and Y have in common
  - Here: a 'stops' tier
  - On this tier, the data looks like:

Attested	tʃ'uspi, rit'i
Unattested	kup'i, k'up'i, kʰp'i

  - Crucially assumes [+constricted glottis] defined only for stops?
- Now, the model can learn a 'tierwise local' constraint on the tier:

Gouskova and Gallagher's results

- A test: non-local restrictions in Quechua (laryngeal), Aymara (laryngeal), Shona (vowels)
- In all three cases, the 'baseline' model does indeed reliably discover *X[seg]Y constraints to induce tiers
- Once the tiers are created, the model learns 'tierwise local' constraints to rule out illegal combinations at a distance
- A cloud on the horizon: baseline models do not always reliably find *X[seg]Y constraints in other languages that they've tried
    - Some technical reasons in how constraint induction is done
    - Some principled reasons: local constraint is 'masked' by other constraints

References

Berent, Iris; Colin Wilson; Gary F. Marcus; and Douglas K. Bemis. 2012. On the role of variables in phonology: Remarks on Hayes and Wilson 2008. Linguistic Inquiry 43, 97–119. URL https://doi.org/10.1162/LING_a_00075. https://direct.mit.edu/ling/article-pdf/43/1/97/724780/ling_a_00075.pdf.

Coetzee, Andries. 2008. Grammaticality and ungrammaticality in phonology. Language 84, 218–257.

Davis, Stuart M. 1984. Some implications of onset-coda constraints for syllable phonology. In Chicago Linguistic Society, volume 20, 46–51.

References (*cont.*)

GOUSKOVA, MARIA and GILLIAN GALLAGHER. 2020. Inducing nonlocal constraints from baseline phonotactics. Natural Language & Linguistic Theory 38, 77–116. URL https://doi.org/10.1007/s11049-019-09446-x.

HAYES, BRUCE and COLIN WILSON. 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry 39, 379–440.