Learning phonotactics of any span and distance

Ignas Rudaitis
ignas.rudaitis@gmail.com
Vilnius University

1 Overview of the contribution

We extend the Multiple Tier-based Strictly 2-Local Inference Algorithm, or MTSL_2IA, of McMullin et al. (2019), in two important respects. (We name the extended version k-MTSLIA.)[1]

Firstly, we relax the previously fixed k-gram span parameter $k = 2$ to arbitrary values ($k \geq 2$). This, for instance, lets us learn[2] the long-distance laryngeal restrictions in South Bolivian Quechua (§3.1; Gallagher, 2010), simultaneous with an allophonic vowel distribution conditioned by consonants (§3.2; Gallagher, 2016). The former is a TSL_2 pattern, and the latter a TSL_3 one, both of which combine into a MTSL_3 pattern by means of intersection.

Secondly, we make the algorithm's restriction against overlapping tiers optional (as defined in Aksënova and Deshmukh, 2018). This does not expand the coverage of attested learnable patterns, nonetheless, it lets us provide the first implemented learner of the definition-true MTSL_k class. The latter class is equal to the intersection closure of TSL_k, as usually defined in the subregular literature (Aksënova and Deshmukh, 2018; Aksënova et al., 2020). Owing to its simpler definition, this version of MTSL_k is easier to manipulate mathematically.

2 Properties of the algorithm

2.1 Running time

Our algorithm, k-MTSLIA, consists of two separate routines $\textsc{learn}(X)$ and $\textsc{scan}(y, G)$. The former constructs a grammar for the sample X, and the latter checks if a newly observed string y conforms to the grammar G. The grammar is returned in an implicit form (§4.2), which makes it possible to run both routines in polynomial time – even when the number of restriction-bearing tiers is bounded only exponentially[3]. $\textsc{Learn}(X)$ runs in $\mathcal{O}(N^k)$, where $N = \sum_{x \in X} |x|$, and $\textsc{scan}(y, G)$ in $\mathcal{O}(|y|^k)$. The degree of the polynomial, k, is the k-gram span parameter. For example, using the algorithm for MTSL_3 entails $k = 3$, and, therefore, cubic running time.

2.2 Minimality of resulting stringsets

For each k-gram $\rho_1 \rho_2 \cdots \rho_k$, k-MTSLIA collects the minimal conditions that a tier τ has to satisfy in order to have $*\rho_1 \rho_2 \cdots \rho_k$ restricted on that tier (§4.2). Therefore, the set of restriction-bearing tiers is necessarily maximal, and, consequently, the stringset accepted by $\textsc{scan}$ is always the minimal MTSL_k superset of the input sample received by $\textsc{learn}$. The fact that TSL_k^τ[4] classes of stringsets are lattice classes (Heinz et al., 2011, 2012) guarantees the uniqueness of such a superset. This also ensures that for each k separately, k-MTSLIA identifies MTSL_k stringsets in the limit – in a *TxtEx* setting – in the terms of Gold (1967).

Additionally, for small alphabets, we have verified the above claim of minimality by comparing k-MTSLIA's results with the outcomes of a brute-force MTSL_k learner.

3 Some phonotactic restrictions of South Bolivian Quechua

3.1 Laryngeal restrictions (TSL_2)

As per Gallagher (2010), South Bolivian Quechua allows only one aspirated *or* ejective stop per word:

kintu	'a bunch'	only plain stops
k'inti	'a pair'	one ejective
k^hastuy	'to chew'	one aspirate
*k'int'i		two ejectives
*$\text{k}^\text{h}\text{ast}^\text{h}$uy		two aspirates

[1] https://github.com/antecedent/k-mtslia

[2] Given that all relevant tier-based trigrams are attested, which is feasible with curated datasets, but less so with naturalistic ones (Wilson and Gallagher, 2018).

[3] As a result of dispensing with the "no overlapping tiers" requirement.

[4] That is, TSL_k with a fixed tier τ.

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 397-399.
Amherst, Massachusetts, June 15-17, 2023

Therefore, on a tier τ_1 containing only stop consonants, one observes the bigram restrictions of the shape *C^{h}C^h, *C^{h}C', *C'C^h, and *C'C'.

3.2 Distribution of mid vowels (TSL$_3$)

As per Gallagher (2016), the same varieties of Quechua exhibit allophonic variation between the high [i, u] and mid [e, o], depending on the presence of uvular stops (Q) nearby. Concretely, [e, o] occur only (1) when there is an uvular stop directly to the left or right, *or* (2) when there is an uvular stop separated from the vowel by a single intervening consonant:

q'epij	'to carry'	(1)
q^{h}eʎu	'lazy'	(1)
erqe	'son'	(2) and (1)
*k'epij		neither (1) nor (2)
*k^{h}elu		neither (1) nor (2)

As a result, [e, o] cannot occur between two non-uvular-stop consonants ($\bar{\text{Q}}$), except when they are non-uvular-stop consonants intervening between the vowel and an uvular stop. By inspecting the Quechua vocabulary included[5] with the Inductive Projection Learner (Gouskova and Gallagher, 2020), we discover that two adjacent stops never occur, therefore, the aforementioned intervening consonants can be clarified as non-stops ($\bar{\text{T}}$).

Avoiding typologically anomalous 5-gram restrictions such as *$\bar{\text{Q}}\bar{\text{T}}^?\text{e}\bar{\text{T}}^?\bar{\text{Q}}$, we instead opt to place an analogous trigram restriction *$\bar{\text{Q}}\text{e}\bar{\text{Q}}$ (and *$\bar{\text{Q}}\text{o}\bar{\text{Q}}$) on the tier $\tau_2 = \bar{\text{Q}} \cup \{\text{e}, \text{o}\}$.

4 Underpinnings of the algorithm

4.1 Paths

k-MTSLIA relies on a similar notion of "paths" as the original MTSL$_2$IA does.

Definition. (k-*path.*) A string x contains a k-path $\langle \rho_1\rho_2 \cdots \rho_k,\, S\rangle$ if and only if all of the following are true:

- x has ρ_1 as its first character,
- x has ρ_k as its last character,
- x has $\rho_1\rho_2 \cdots \rho_k$ as one of its subsequences,
- x has only one such subsequence,
- erasing this subsequence from x leaves x',
- S is the set of (distinct) characters in x', *and*
- S and $\{\rho_1, \rho_2, \ldots, \rho_k\}$ are disjoint.

[5] https://github.com/gouskova/inductive_projection_learner/tree/master/data/quechua

For instance, "q^{h}eʎu" possesses the 3-paths $\langle$q^{h}ʎu, $\{$e$\}\rangle$ and $\langle$q^{h}eu, $\{$ʎ$\}\rangle$.

4.2 Interpretation of attested paths

Each time we witness a k-path $\langle \rho_1\rho_2 \cdots \rho_k,\, S\rangle$ in the inputs (including substrings), we can restrict $*\rho_1\rho_2 \cdots \rho_k$ on some tier τ, but we must have at least one extra character from S on the tier. Only this way will the k-gram $\rho_1\rho_2 \cdots \rho_k$ be broken apart (by the intervening character) in the tier image of some input string. If it were not broken apart in this manner, the entire k-gram would project onto the tier image and make the restriction $*\rho_1\rho_2 \cdots \rho_k$ contradictory with our data.

To put it differently, each k-path $\langle \rho_1\rho_2 \cdots \rho_k,\, \{\sigma_1, \sigma_2, \ldots, \sigma_N\}\rangle$ can be interpreted as follows:

$$\sigma_1 \in \tau \;\vee\; \sigma_2 \in \tau \;\vee\; \cdots \;\vee\; \sigma_N \in \tau.$$

Consider a certain 3-path of the string "q'epij", namely, $\langle$q'pj, $\{$e, i$\}\rangle$. In k-MTSLIA, it will be interpreted in the following way:

$$\text{e} \in \tau \;\vee\; \text{i} \in \tau.$$

In isolation, this formula would entail that *q'pj will be restricted on the tiers $\{$q', p, j, e$\}$, $\{$q', p, j, i$\}$, and $\{$q', p, j, e, i$\}$ – that is, on all tiers τ that satisfy the formula (and contain the k-gram itself).

These disjunctive clauses make up the grammars that k-MTSLIA's LEARN routine returns. In fact, each attested k-path contributes to such a disjunctive clause, all of which are eventually conjoined in one CNF formula that constitutes the grammar itself – or, more precisely, it constitutes one portion of it, associated with a specific k-gram restriction.

The SCAN(y) routine then checks whether the conjunction $G \wedge \neg G'$ of a given grammar G and the negation of another grammar $G' = \text{LEARN}(\{y\})$ is a satisfiable formula. If it is, the string y is rejected. This is a simple procedure, linear in the size of the conjoined formula, owing to the fact that the grammars only contain single-polarity literals.

5 Discussion

We hope to have enriched the toolset of subregular grammatical inference with a polynomial-time algorithm for a known learning problem. However, while potentially useful for computational experiment-heavy research on the topic, k-MTSLIA generalizes too little to be a cognitively realistic learner of phonotactics for $k \geq 3$ – unless an additional source of inductive bias is provided.

References

Alëna Aksënova and Sanket Deshmukh. 2018. Formal restrictions on multiple tiers. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 64–73.

Alëna Aksënova, Jonathan Rawski, Thomas Graf, and Jeffrey Heinz. 2020. The computational power of harmony.

Gillian Gallagher. 2010. *The perceptual basis of long-distance laryngeal restrictions*. Ph.D. thesis, Massachusetts Institute of Technology.

Gillian Gallagher. 2016. Vowel height allophony and dorsal place contrasts in Cochabamba Quechua. *Phonetica*, 73(2):101–119.

E Mark Gold. 1967. Language identification in the limit. *Information and control*, 10(5):447–474.

Maria Gouskova and Gillian Gallagher. 2020. Inducing nonlocal constraints from baseline phonotactics. *Natural Language & Linguistic Theory*, 38:77–116.

Jeffrey Heinz, Anna Kasprzik, and Timo Kötzing. 2012. Learning in the limit with lattice-structured hypothesis spaces. *Theoretical Computer Science*, 457:111–127.

Jeffrey Heinz, Chetan Rawal, and Herbert G Tanner. 2011. Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human language technologies*, pages 58–64.

Kevin McMullin, Alëna Aksënova, and Aniello De Santo. 2019. Learning phonotactic restrictions on multiple tiers. *Proceedings of the Society for Computation in Linguistics*, 2(1):377–378.

Colin Wilson and Gillian Gallagher. 2018. Accidental gaps and surface-based phonotactic learning: A case study of South Bolivian Quechua. *Linguistic Inquiry*, 49(3):610–623.