Learning phonotactics of any span and distance

Ignas Rudaitis

Vilnius University, ignas.rudaitis@gmail.com
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 MTSL2IA, of McMullin et al. (2019), in two important respects. (We name the extended version $k$-MTSLIA.)

Firstly, we relax the previously fixed $k$-gram span parameter $k = 2$ to arbitrary values ($k \geq 2$). This, for instance, lets us learn 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 Aksenova 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 (Aksenova and Deshmukh, 2018; Aksenova 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 LEARN($X$) and 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$^3$. LEARN($X$) runs in $O(N^k)$, where $N = \sum_{x \in X} |x|$, and SCAN($y$, $G$) in $O(|y|^k)$. 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 $\tau$ 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 SCAN is always the minimal MTSL$_k$ superset of the input sample received by LEARN. The fact that TSL$_{k,4}^2$ 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$^h$astuy ‘to chew’ one aspirate
- *k’inti two ejectives
- *k$^h$astuy two aspirates

$^3$As a result of dispensing with the “no overlapping tiers” requirement.

$^4$That is, TSL$_k$ with a fixed tier $\tau$. 

Amherst, Massachusetts, June 15-17, 2023
Therefore, on a tier $\tau_1$ containing only stop consonants, one observes the bigram restrictions of the shape $*C'hCh$, $*C'Ch$, $*C'C'$, 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:

- $q^he\ddot{u}$ ‘to carry’ (1)
- $q^he\ddot{u}$ ‘lazy’ (1)
- erqe ‘son’ (2) and (1)
- *$ke'$pij neither (1) nor (2)
- *$khe'u$ neither (1) nor (2)

As a result, [e, o] cannot occur between two non-uvular-stop consonants (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 (T).

Avoiding typologically anomalous 5-gram restrictions such as $*Q'T'e'T'Q$, we instead opt to place an analogous trigram restriction $*QeQ$ (and $*QoQ$) on the tier $\tau_2 = Q \cup \{e, o\}$.

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 $\rho_1$ as its first character,
- $x$ has $\rho_k$ 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.

For instance, “$q^he\ddot{u}$” possesses the 3-paths $\langle q^he\ddot{u}, \{e\} \rangle$ and $\langle q^he\ddot{u}, \{A\} \rangle$.

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 $\tau$, 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:

$$\sigma_1 \in \tau \lor \sigma_2 \in \tau \lor \cdots \lor \sigma_N \in \tau.$$

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:

$$e \in \tau \lor i \in \tau.$$

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 $\tau$ 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 $G \land \neg G'$ 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


