Syntax Repair as Language Intersection

ANONYMOUS AUTHOR(S)

We introduce a new technique for correcting syntax errors in arbitrary context-free languages. Our work addresses the problem of syntax error correction, which we solve by defining a finite language that provably generates every repair within a certain edit distance. To do this, we adapt the Bar-Hillel construction from formal languages, guaranteeing this language is sound and complete with respect to a programming language's grammar. This technique also admits a polylogarithmic time algorithm for deciding intersection nonemptiness between CFLs and acyclic NFAs, the first of its kind in the parsing literature.

1 INTRODUCTION

When programming, one invariably encounters a recurring scenario in which the editor occupies an unparseable state. Faced with this predicament, programmers must spend time to locate and repair the error before proceeding. We solve this problem automatically by generating a list of candidate repairs which contains with high probability the true repair, assuming this repair differs by no more than a few edits.

Prior research in syntax repair uses stochastic and deterministic methods. In the case of neural program repair, the model is typically used as a base distribution, however most neural language models are only approximate inference methods and thus require some form of postprocessing and rejection sampling to ensure the results conform to the grammar. In contrast, most rule-based syntax repair methods, while exact, are too weak to fully model the full distribution of natural source code and must rely on local heuristics.

The primary issue with both of these methods is that they generate a very low number of repairs. Even if the repair model is highly accurate, it is unlikely to generate the true repair when there are a large number of approximately equidistant or equiprobable repairs.

To address this problem, our method explicitly generates and ranks every repair within a fixed edit distance. This ensures that the true repair is retrieved and scored, as long as it is only a few edits apart. Specifically, we employ the Bar-Hillel construction from formal language theory to generate a regular expression representing the finite language, then enumerate up to 2^{16} distinct words from this language, rerank the enumerated set using a neural language model, and present the top dozen or so results to the user, all within a few hundred milliseconds.

We evaluate our approach on a dataset of human syntax errors and fixes fewer than five lexical edits apart and shorter than 120 tokens, large enough to fit a few lines of source code in realistic programming languages and editing scenarios. Our work shows this technique is extremely effective at predicting the true repair across a dataset of Python source code, on average 5x more accurately than the previous state of the art at the same latency.

2 BACKGROUND

Recall that a CFG, $\mathcal{G} = \langle \Sigma, V, P, S \rangle$, is a quadruple consisting of terminals (Σ), nonterminals (V), productions ($P \colon V \to (V \mid \Sigma)^+$), and a start symbol, ($S$). Every CFG is reducible to so-called *Chomsky Normal Form* [2], $P' \colon V \to (V^2 \mid \Sigma)$, where every production is either (1) a binary production $w \to xz$, or (2) a unit production $w \to t$, where $w, x, z : V$ and $t : \Sigma$. For example:

$$G = \{\, S \to S\,S \mid (\,S\,) \mid (\,) \,\} \Longrightarrow G' = \{\, S \to Q\,R \mid S\,S \mid L\,R, \quad R \to), \quad L \to (, \quad Q \to L\,S \,\}$$

Likewise, a finite state automaton (FSA) is a quintuple $\mathcal{A} = \langle Q, \Sigma, \delta, I, F \rangle$, where Q is a finite set of states, Σ is a finite alphabet, $\delta \subseteq Q \times \Sigma \times Q$ is the transition function, and $I, F \subseteq Q$ are the set of initial and final states, respectively. We will adhere to this notation in the following sections.

There is an equivalent characterization of the regular languages using an inductively defined datatype which is often more elegant to work with. Consider the generalized regular expression (GRE) fragment containing concatenation, conjunction and disjunction:

Definition 2.1 (Generalized Regex). Let e be an expression defined by the grammar:

$$e ::= \varnothing \mid \varepsilon \mid \Sigma \mid e \cdot e \mid e \vee e \mid e \wedge e$$

Semantically, we can interpret these expressions as denoting regular languages:

$$\begin{aligned} \mathcal{L}(\quad \varnothing \quad) &= \varnothing & \qquad \mathcal{L}(\quad x \cdot z \quad) &= \mathcal{L}(x) \times \mathcal{L}(z)^{1} \\ \mathcal{L}(\quad \varepsilon \quad) &= \{\varepsilon\} & \mathcal{L}(\quad x \vee z \quad) &= \mathcal{L}(x) \cup \mathcal{L}(z) \\ \mathcal{L}(\quad a \quad) &= \{a\} & \mathcal{L}(\quad x \wedge z \quad) &= \mathcal{L}(x) \cap \mathcal{L}(z) \end{aligned}$$

Brzozowski [1] introduces the concept of differentiation, which allows us to quotient a regular language by some given prefix.

Definition 2.2 (Brzozowski, 1964). To compute the quotient $\partial_a(L) = \{b \mid ab \in L\}$, we:

$$\begin{aligned} \partial_a(\quad \varnothing \quad) &= \varnothing & \qquad \delta(\quad \varnothing \quad) &= \varnothing \\ \partial_a(\quad \varepsilon \quad) &= \varnothing & \delta(\quad \varepsilon \quad) &= \varepsilon \\ \partial_a(\quad b \quad) &= \begin{cases} \varepsilon & \text{if } a = b \\ \varnothing & \text{if } a \neq b \end{cases} & \delta(\quad a \quad) &= \varnothing \\ \partial_a(\quad x \cdot z \quad) &= (\partial_a x) \cdot z \vee \delta(x) \cdot \partial_a z & \delta(\quad x \cdot z \quad) &= \delta(x) \wedge \delta(z) \\ \partial_a(\quad x \vee z \quad) &= \partial_a x \vee \partial_a z & \delta(\quad x \vee z \quad) &= \delta(x) \vee \delta(z) \\ \partial_a(\quad x \wedge z \quad) &= \partial_a x \wedge \partial_a z & \delta(\quad x \wedge z \quad) &= \delta(x) \wedge \delta(z) \end{aligned}$$

Primarily, this gadget was designed to handle membership queries, for which purpose it has received considerable attention in recent years:

Theorem 2.3 (Recognition). *For any regex e and $\sigma : \Sigma^*$, $\sigma \in \mathcal{L}(e) \iff \varepsilon \in \mathcal{L}(\partial_\sigma e)$, where:*

$$\partial_\sigma(e) : E \to E = \begin{cases} e & \text{if } \sigma = \varepsilon \\ \partial_b(\partial_a e) & \text{if } \sigma = a \cdot b, a \in \Sigma, b \in \Sigma^* \end{cases}$$

Variations on this basic procedure can also be used for functional parsing and regular expression tasks. Brzozowski's derivative can also be used to decode witnesses. We will first focus on the nonempty disjunctive fragment, and define this process in two steps:

[1] Or $\{a \cdot b \mid a \in \mathcal{L}(x) \wedge b \in \mathcal{L}(z)\}$ to be more precise, however we make no distinction.

Theorem 2.4 (Generation). *For any nonempty $(\varepsilon, \wedge)$-free regex, e, to witness $\sigma \in \mathcal{L}(e)$:*

$$\texttt{follow}(e) : E \longrightarrow 2^{\Sigma} = \begin{cases} \{e\} & \text{if } e \in \Sigma \\ \texttt{follow}(x) & \text{if } e = x \cdot z \\ \texttt{follow}(x) \cup \texttt{follow}(z) & \text{if } e = x \vee z \end{cases}$$

$$\texttt{choose}(e) : E \longrightarrow \Sigma^{+} = \begin{cases} e & \text{if } e \in \Sigma \\ \left(s \xleftarrow{\$} \texttt{follow}(e)\right) \cdot \texttt{choose}(\partial_s e) & \text{if } e = x \cdot z \\ \texttt{choose}(e' \xleftarrow{\$} \{x, z\}) & \text{if } e = x \vee z \end{cases}$$

Here, we use the $\xleftarrow{\$}$ operator to denote probabilistic choice, however any deterministic choice function will also suffice to generate a witness. Now we are equipped to handle conjunction.

2.1 Language intersection

We will now define intersection in a slightly more expressive manner, which has the added benefit of being more readily parallelizable. Recall every regular language is context-free. Therefor, to take the intersection between two regular languages, we can treat one as a CFL, which often admits a more compact representation. Alternatively, we can take the intersection between a truly non-regular CFL (such as a programming language syntax) and some regular language.

Theorem 2.5 (Bar-Hillel, 1961). *For any context-free grammar (CFG), $G = \langle V, \Sigma, P, S\rangle$, and nondeterministic finite automata, $A = \langle Q, \Sigma, \delta, I, F\rangle$, there exists a CFG $G_\cap = \langle V_\cap, \Sigma_\cap, P_\cap, S_\cap\rangle$ such that $\mathcal{L}(G_\cap) = \mathcal{L}(G) \cap \mathcal{L}(A)$.*

Salomaa [5] introduces a direct, but inefficient construction:

Definition 2.6 (Salomaa, 1973). One could construct $G_\cap$ like so,

$$\frac{q \in I \quad r \in F}{(S \to qSr) \in P_\cap}\ \surd \qquad \frac{(w \to a) \in P \qquad (q \xrightarrow{a} r) \in \delta}{(qwr \to a) \in P_\cap}\ \uparrow \qquad \frac{(w \to xz) \in P \qquad p, q, r \in Q}{(pwr \to (pxq)(qzr)) \in P_\cap}\ \bowtie$$

however most synthetic productions in $P_\cap$ will be non-generating or unreachable. This naïve method will construct a synthetic production for state pairs which are not even connected by any path, which is clearly excessive.

3 INFORMAL STATEMENT

Assume there exists a transducer from Unicode tokens to grammatical tokens, $\tau : \Sigma_U^* \to \Sigma_G^*$. In the compiler nomenclature τ is called a *lexer* and would typically be regular under mild assumptions. Lexing abstracts over named identifiers and primitively typed literals, and removes whitespace or otherwise replaces them with an alphanumeric identifier in whitespace-sensitive languages. In this paper, we do not consider τ and strictly deal with languages over Σ_G^*, or simply Σ^* for brevity.

Now suppose we have a syntax, $\ell \subset \Sigma^*$, containing every acceptable program. A syntax error is an unacceptable string, $\underset{\sim}{\sigma} \notin \ell$, that we wish to repair. We can model syntax repair as a language intersection between a context-free language (CFL) and a regular language. Henceforth, $\underset{\sim}{\sigma}$ will always and only be used to denote a syntactically invalid string whose target language is known.

Given the source code for a computer program $\underset{\sim}{\sigma}$ and a grammar G, our goal is to find every valid string σ consistent with the grammar G and within a certain edit distance, d. Consider the language of valid strings within a given Levenshtein distance from a reference string $\underset{\sim}{\sigma}$. We can intersect the language given by the Levenshtein automaton with the language of all valid programs given by the grammar G. The resulting language $\mathcal{L}(G_\cap)$ will contain every repair within the designated edit distance.

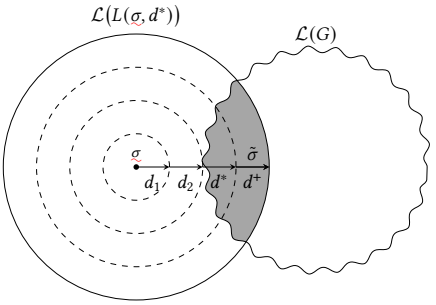


Fig. 1. CFL intersection with the local edit region of a given broken code snippet.

Once the repairs are retrieved, they can be mapped back into raw source code tokens in a straightforward manner. The core algorithm is formatting-oblivious. After the repairs are returned, we compute an edit alignment over lexical tokens, restore unmodified tokens to their Unicode form, add placeholders for fresh names, numbers, and string literals, then finally apply an off-the-shelf code formatter to return the concrete suggestions. Both the preprocessing and the formatting steps are tangential to this paper, in which we confine ourselves to a lexical alphabet.

4 FORMAL STATEMENT

Definition 4.1 (Bounded Levenshtein-CFL reachability). Given a CFL, ℓ, and an invalid string, $\underset{\sim}{\sigma} : \bar{\ell}$, find every valid string reachable within d edits of $\underset{\sim}{\sigma}$, i.e., letting Δ be the Levenshtein metric and $L(\underset{\sim}{\sigma}, d) = \{\sigma' \mid \Delta(\underset{\sim}{\sigma}, \sigma') \leq d\}$ be the Levenshtein d-ball, we seek to find $\ell_\cap = L(\underset{\sim}{\sigma}, d) \cap \ell$.

As the admissible set $\ell_\cap$ is typically under-constrained, we want a procedure which surfaces natural and valid repairs over unnatural but valid repairs:

Definition 4.2 (Ranked repair). Given a finite language $\ell_\cap = L(\underset{\sim}{\sigma}, d) \cap \ell$ and a probabilistic language model $\mathrm{P}_\theta : \Sigma^* \to [0, 1] \subset \mathbb{R}$, the ranked repair problem is to find the top-k maximum likelihood repairs under the language model. That is,

$$R(A, P_\theta) = \underset{\boldsymbol{\sigma} \subseteq \ell_\cap, |\boldsymbol{\sigma}| \leq k}{\operatorname{argmax}} \sum_{\sigma \in \boldsymbol{\sigma}} \mathrm{P}_\theta(\sigma) \tag{1}$$

A popular approach to ranked repair involves learning a distribution over strings, however this is highly sample-inefficient and generalizes poorly to new languages. Approximating a distribution over Σ^* forces the model to jointly learn syntax and stylometry. Furthermore, even with an extremely efficient approximate sampler for $\sigma \sim \ell_\cap$, due to the size of ℓ and $L(\underset{\sim}{\sigma}, d)$, it would be intractable to sample either ℓ or $L(\underset{\sim}{\sigma}, d)$, reject duplicates, then reject invalid ($\sigma \notin \ell$) or unreachable $\left(\sigma \notin L(\underset{\sim}{\sigma}, d)\right)$ edits, and completely out of the question to sample $\sigma \sim \Sigma^*$ as do many neural language models.

As we will demonstrate, the ranked repair problem can be factorized into a bilevel objective: first maximal retrieval, then ranking. Instead of working with strings, we will explicitly construct a grammar which soundly and completely generates the set $\ell \cap L(\underset{\sim}{\sigma}, d)$, then retrieve repairs from its language. By ensuring retrieval is sufficiently precise and exhaustive, maximizing likelihood over the retrieved set can be achieved with a much simpler, syntax-oblivious language model.

Assuming we have a grammar that recognizes the Levenshtein-CFL intersection, the question then becomes how to maximize the number of unique valid sentences in a given number of samples. Top-down incremental sampling with replacement eventually converges to the language, but does so superlinearly [**?**]. Due to practical considerations including latency, we require the sampler to converge linearly, ensuring with much higher probability that natural repairs are retrieved in a timely manner. This motivates the need for a specialized generating function. More precisely,

Definition 4.3 (Linear convergence). Given a finite CFL, ℓ, we want a randomized generating function, $\boldsymbol{\varphi} : \mathbb{N}_{\leq |\ell|} \to 2^\ell$, whose rate of convergence is linear in expectation, i.e., $\mathbb{E}_{i \in [1,n]} |\boldsymbol{\varphi}(i)| \propto n$.

This will ensure that if $|\ell_\cap|$ is sufficiently small and enough samples are drawn, $\boldsymbol{\varphi}$ is sure to include a representative subset, and additionally, will terminate after exhausting all valid repairs.

To satisfy Def. 4.3, we can construct a bijection from syntax trees to integers (§ **??**), sample integers uniformly without replacement, then decode them as trees. This will produce a set of unique trees, and each tree, assuming grammatical unambiguity, will correspond to a unique sentence in the language. Finally, sentences can be scored and ranked by likelihood under a language model.

Otherwise, if the grammar, G_ℓ, is ambiguous, it can be translated into a DFA, then decoded (§ **??**) using an autoregressive language model or any suitably fast scoring function of the implementer's choice. In our case, we use a low-order Markov model for its inference speed, data efficiency, and simplicity. So long as the decoder samples ℓ without replacement, it will satisfy Def. 4.3.

5 METHOD

Our method is to treat finite language intersections as matrix exponentiation.

Theorem 5.1. *For every CFG, G, and every acyclic NFA (ANFA), A, there exists a decision procedure* $\varphi : CFG \to ANFA \to \mathbb{B}$ *such that* $\varphi(G, A) \models [\mathcal{L}(G) \cap \mathcal{L}(A) \neq \varnothing]$ *which requires* $\mathcal{O}\big((\log |Q|)^c\big)$ *time using* $\mathcal{O}\big((|V||Q|)^k\big)$ *parallel processors for some* $c, k < \infty$.

Proof. WTS there exists a path $p \rightsquigarrow r$ in A such that $p \in I, r \in F$ where $p \rightsquigarrow r \vdash S$.

There are two cases, at least one of which must hold for $w \in V$ to parse a given $p \rightsquigarrow r$ pair:

(1) p steps directly to r in which case it suffices to check $\exists a.\big((p \xrightarrow{a} r) \in \delta \wedge (w \to a) \in P\big)$, or,

(2) there is some midpoint $q \in Q$, $p \rightsquigarrow q \rightsquigarrow r$ such that $\exists x, z.\big((w \to xz) \in P \wedge \overbrace{\underbrace{p \rightsquigarrow q}_{x}, \underbrace{q \rightsquigarrow r}_{z}}^{w}\big)$.

This decomposition immediately suggests a dynamic programming solution. Let M be a matrix of type $E^{|Q|\times|Q|\times|V|}$ indexed by Q. Since we assumed δ is acyclic, there exists a topological sort of δ imposing a total order on Q such that M is strictly upper triangular (SUT). Initiate it thusly:

$$M_0[r, c, w] = \bigvee_{a \in \Sigma} \{a \mid (w \to a) \in P \wedge (q_r \xrightarrow{a} q_c) \in \delta\} \tag{2}$$

The algebraic operations $\oplus, \otimes : E^{2|V|} \to E^{|V|}$ will be defined elementwise:

$$[\ell \oplus r]_w = [\ell_w \vee r_w] \tag{3}$$

$$[\ell \otimes r]_w = \bigvee_{x,z \in V} \{\ell_x \cdot r_z \mid (w \to xz) \in P\} \tag{4}$$

By slight abuse of notation[2], we will redefine the matrix exponential over this domain as:

$$\exp(M) = \sum_{i=0}^{\infty} M_0^i = \sum_{i=0}^{|Q|} M_0^i \text{ (since } M \text{ is SUT.)} \tag{5}$$

To solve for the fixpoint, we can instead use exponentiation by squaring:

$$T(2n) = \begin{cases} M_0, & \text{if } n = 1, \\ T(n) + T(n)^2 & \text{otherwise.} \end{cases} \tag{6}$$

Therefor, we only need at most $\lceil \log_2 |Q| \rceil$ sequential steps to reach the fixpoint. Finally, we will union all the languages of every state pair deriving S into a new nonterminal, $S_\cap$.

$$S_\cap = \bigvee_{q \in I,\, q' \in F} \exp(M)[q, q', S] \text{ and } \varphi = [S_\cap \neq \varnothing] \tag{7}$$

To decode a witness in case of non-emptiness, one may simply `choose`$(S_\cap)$. □

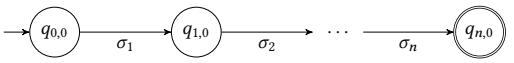
[2]Customarily, there is a $\frac{1}{k!}$ factor to suppress exploding entries, but alas this domain has no multiplicative inverse.

6 EXAMPLES

In this section, we will consider three examples of intersections with finite languages. First, parsing can be viewed as a special case of language intersection with an automaton accepting a single word. Second, completion can be seen as a case of intersection with terminal wildcards in known locations. Thirdly, we consider syntax repair, where we will intersect a language representing all possible edit paths to determine the edit location(s) and fill them with appropriate terminals.

6.1 Recognition as intersection

In the case of ordinary CFL recognition, the automaton accepts just a single word:



Given a CFG, $G' : \mathcal{G}$ in Chomsky Normal Form (CNF), we can construct a recognizer $R : \mathcal{G} \to \Sigma^n \to \mathbb{B}$ for strings $\sigma : \Sigma^n$ as follows. Let 2^V be our domain, 0 be $\varnothing$, $\oplus$ be $\cup$, and $\otimes$ be defined as:

$$X \otimes Z = \left\{ w \mid \langle x, z \rangle \in X \times Z, (w \to xz) \in P \right\} \tag{8}$$

If we define $\hat{\sigma}_r = \{w \mid (w \to \sigma_r) \in P\}$, then construct a matrix with nonterminals on the superdiagonal representing each token, $M_0[r+1 = c](G', \sigma) = \hat{\sigma}_r$, the fixpoint $M_{i+1} = M_i + M_i^2$ is uniquely determined by the superdiagonal entries. Omitting the exponentiation-by-squaring detail, the ordinary fixedpoint iteration simply fills successive diagonals:

$$M_0 = \begin{pmatrix} \varnothing & \hat{\sigma}_1 & \varnothing & \cdots & \varnothing \\ & \ddots & \ddots & & \\ & & & & \varnothing \\ & & & & \hat{\sigma}_n \\ \varnothing & & & & \varnothing \end{pmatrix} \Rightarrow \begin{pmatrix} \varnothing & \hat{\sigma}_1 & \Lambda & \cdots & \varnothing \\ & \ddots & \ddots & & \\ & & & & \Lambda \\ & & & & \hat{\sigma}_n \\ \varnothing & & & & \varnothing \end{pmatrix} \Rightarrow \ldots \Rightarrow M_\infty = \begin{pmatrix} \varnothing & \hat{\sigma}_1 & \Lambda & \cdots & \Lambda^*_\sigma \\ & \ddots & \ddots & & \\ & & & & \Lambda \\ & & & & \hat{\sigma}_n \\ \varnothing & & & & \varnothing \end{pmatrix}$$

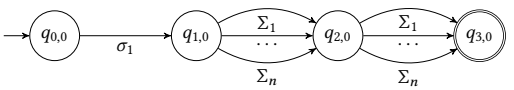
Once the fixpoint M_∞ is attained, the proposition $[S \in \Lambda^*_\sigma]$ decides language membership, i.e., $[\sigma \in \mathcal{L}(G)]$ [3]. So far, this procedure is essentially the textbook CYK algorithm in a linear algebraic notation [3] and a well-established technique in the parsing literature [4].

6.2 Completion as intersection

We can also consider a more general automaton for completing a string with holes, representing edits in fixed locations which can be filled by any terminal, which we call *completion*. In this case, the fixpoint is characterized by a system of language equations, whose solutions are the set of all sentences consistent with the template.

Definition 6.1 (Completion). Let $\underline{\Sigma} = \Sigma \cup \{\_\}$, where _ denotes a hole. We denote $\sqsubseteq : \Sigma^n \times \underline{\Sigma}^n$ as the relation $\{\langle \sigma', \sigma \rangle \mid \sigma_i \in \Sigma \implies \sigma'_i = \sigma_i\}$ and the set of all inhabitants $\{\sigma' : \Sigma^+ \mid \sigma' \sqsubseteq \sigma\}$ as $\mathrm{H}(\sigma)$. Given a *porous string*, $\sigma : \underline{\Sigma}^*$ we seek all syntactically valid inhabitants, i.e., $A(\sigma) = \mathrm{H}(\sigma) \cap \ell$.

Here, the FSA takes a similar shape but can have multiple arcs between subsequent states, e.g.:



[3] Hereinafter, we use Iverson brackets to denote the indicator function of a predicate with free variables, i.e., $[P] \Leftrightarrow \mathbb{1}(P)$.

This corresponds to a template with two holes, $\sigma = 1\ \_\ \_$. Suppose the context-free grammar is $G = \{S \to NON, O \to + \mid \times, N \to 0 \mid 1\}$. This grammar will first be rewritten into CNF as $G' = \{S \to NL, N \to 0 \mid 1, O \to \times \mid +, L \to ON\}$. Using the powerset algebra we just defined, the matrix fixpoint $M' = M + M^2$ can be computed as follows, shown in the leftmost column below:

	2^V	$\mathbb{Z}_2^{\lvert V\rvert}$	$\mathbb{Z}_2^{\lvert V\rvert} \to \mathbb{Z}_2^{\lvert V\rvert}$
M_0	$\begin{pmatrix} \{N\} & & \\ & \{N,O\} & \\ & & \{N,O\} \end{pmatrix}$	$\begin{pmatrix} L\,N\,O\,S \\ \square\blacksquare\square\square & & \\ & \square\blacksquare\blacksquare\square & \\ & & \square\blacksquare\blacksquare\square \end{pmatrix}$	$\begin{pmatrix} V_{0,1} & & \\ & V_{1,2} & \\ & & V_{2,3} \end{pmatrix}$
M_1	$\begin{pmatrix} \{N\} & \varnothing & \\ & \{N,O\} & \{L\} \\ & & \{N,O\} \end{pmatrix}$	$\begin{pmatrix} \square\blacksquare\square\square & \square\square\square\square & \\ & \square\blacksquare\blacksquare\square & \blacksquare\square\square\square \\ & & \square\blacksquare\blacksquare\square \end{pmatrix}$	$\begin{pmatrix} V_{0,1} & V_{0,2} & \\ & V_{1,2} & V_{1,3} \\ & & V_{2,3} \end{pmatrix}$
$M_2 = M_\infty$	$\begin{pmatrix} \{N\} & \varnothing & \{S\} \\ & \{N,O\} & \{L\} \\ & & \{N,O\} \end{pmatrix}$	$\begin{pmatrix} \square\blacksquare\square\square & \square\square\square\square & \square\square\square\blacksquare \\ & \square\blacksquare\blacksquare\square & \blacksquare\square\square\square \\ & & \square\blacksquare\blacksquare\square \end{pmatrix}$	$\begin{pmatrix} V_{0,1} & V_{0,2} & V_{0,3} \\ & V_{1,2} & V_{1,3} \\ & & V_{2,3} \end{pmatrix}$

The same procedure can be translated, without loss of generality, into the bit domain ($\mathbb{Z}_2^{|V|}$) using a lexicographic nonterminal ordering, however M_∞ in both 2^V and $\mathbb{Z}_2^{|V|}$ represents a decision procedure, i.e., $[S \in V_{0,3}] \Leftrightarrow [V_{0,3,3} = \blacksquare] \Leftrightarrow [A(\sigma) \neq \varnothing]$. Since $V_{0,3} = \{S\}$, we know there exists at least one solution $\sigma' \in A(\sigma)$, but M_∞ does not explicitly reveal its identity.

To extract the inhabitants, we can translate the bitwise procedure into an equation with free variables. Here, we can encode the idempotency constraint directly as $M = M^2$. We first define $X \boxtimes Z = [X_2 \wedge Z_1, \bot, \bot, X_1 \wedge Z_0]$ and $X \boxplus Z = [X_i \vee Z_i]_{i \in [0,|V|)}$, mirroring $\oplus, \otimes$ from the powerset domain, now over bitvectors. Since the unit nonterminals O, N can only occur on the superdiagonal, they may be safely ignored by $\boxtimes$. To solve for M_∞, we proceed by first computing $V_{0,2}, V_{1,3}$:

$$\begin{aligned} V_{0,2} &= V_{0,j} \cdot V_{j,2} = V_{0,1} \boxtimes V_{1,2} \\ &= [L \in V_{0,2}, \bot, \bot, S \in V_{0,2}] \\ &= [O \in V_{0,1} \wedge N \in V_{1,2}, \bot, \bot, N \in V_{0,1} \wedge L \in V_{1,2}] \\ &= [V_{0,1,2} \wedge V_{1,2,1}, \bot, \bot, V_{0,1,1} \wedge V_{1,2,0}] \end{aligned}$$

$$\begin{aligned} V_{1,3} &= V_{1,j} \cdot V_{j,3} = V_{1,2} \boxtimes V_{2,3} \\ &= [L \in V_{1,3}, \bot, \bot, S \in V_{1,3}] \\ &= [O \in V_{1,2} \wedge N \in V_{2,3}, \bot, \bot, N \in V_{1,2} \wedge L \in V_{2,3}] \\ &= [V_{1,2,2} \wedge V_{2,3,1}, \bot, \bot, V_{1,2,1} \wedge V_{2,3,0}] \end{aligned}$$

Now we solve for the corner entry $V_{0,3}$ by dotting the first row and last column, which yields:

$$\begin{aligned} V_{0,3} &= V_{0,j} \cdot V_{j,3} = (V_{0,1} \boxtimes V_{1,3}) \boxplus (V_{0,2} \boxtimes V_{2,3}) \\ &= [V_{0,1,2} \wedge V_{1,3,1} \vee V_{0,2,2} \wedge V_{2,3,1}, \bot, \bot, V_{0,1,1} \wedge V_{1,3,0} \vee V_{0,2,1} \wedge V_{2,3,0}] \end{aligned}$$

Since we only care about $V_{0,3,3} \Leftrightarrow [S \in V_{0,3}]$, we can ignore the first three entries and solve for:

$$\begin{aligned} V_{0,3,3} &= V_{0,1,1} \wedge V_{1,3,0} \vee V_{0,2,1} \wedge V_{2,3,0} \\ &= V_{0,1,1} \wedge (V_{1,2,2} \wedge V_{2,3,1}) \vee V_{0,2,1} \wedge \bot \\ &= V_{0,1,1} \wedge V_{1,2,2} \wedge V_{2,3,1} \\ &= [N \in V_{0,1}] \wedge [O \in V_{1,2}] \wedge [N \in V_{2,3}] \end{aligned}$$

6.3 Repair as intersection

Finally, we are ready to consider the general case of syntax repair, in which case the edit locations are not localized but can occur anywhere in the string. In this case, we construct a lattice of all possible edit paths within a fixed distance. This structure is called a Levenshtein automaton.

As the original construction defined by Schultz and Mihov [6] contains cycles and ε-transitions, we propose a variant which is ε-free and acyclic. Furthermore, we adopt a nominal form which supports infinite alphabets and considerably simplifies the language intersection to follow. Illustrated in Fig. 2 is an example of a small Levenshtein automaton recognizing $L(\sigma : \Sigma^5, 3)$. Unlabeled arcs accept any terminal from the alphabet, Σ. Equivalently, this transition system can be viewed as a kind of proof system within an unlabeled lattice. The following construction is equivalent to Schultz and Mihov's original Levenshtein automaton, but is more amenable to our purposes as it does not any contain ε-arcs, and instead uses skip connections to recognize consecutive deletions of varying lengths.

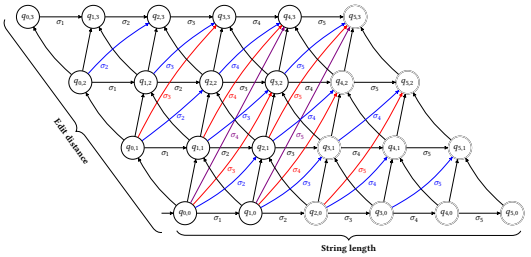


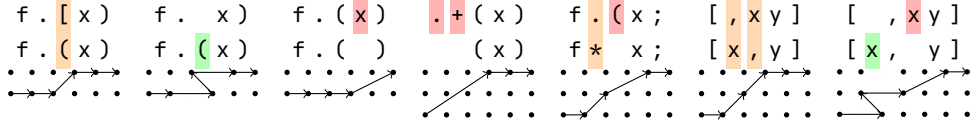
Fig. 2. NFA recognizing Levenshtein $L(\sigma : \Sigma^5, 3)$.

$$\frac{s \in \Sigma \quad i \in [0,n] \quad j \in [1,d_{\max}]}{(q_{i,j-1} \xrightarrow{s} q_{i,j}) \in \delta} \qquad \frac{s \in \Sigma \quad i \in [1,n] \quad j \in [1,d_{\max}]}{(q_{i-1,j-1} \xrightarrow{s} q_{i,j}) \in \delta}$$

$$\frac{i \in [1,n] \quad j \in [0,d_{\max}]}{(q_{i-1,j} \xrightarrow{\sigma_i} q_{i,j}) \in \delta} \qquad \frac{d \in [1,d_{\max}] \quad i \in [d+1,n] \quad j \in [d,d_{\max}]}{(q_{i-d-1,j-d} \xrightarrow{\sigma_i} q_{i,j}) \in \delta}$$

$$\frac{}{q_{0,0} \in I}\ \textsc{Init} \qquad \frac{q_{i,j} \in Q \qquad |n-i+j| \le d_{\max}}{q_{i,j} \in F}\ \textsc{Done}$$

Each arc plays a specific role. ↖ handles insertions, ↗ handles substitutions and ⋰ handles deletions of one or more terminals. Let us consider some illustrative cases.



Note that the same patch can have multiple Levenshtein alignments. Done constructs the final states, which are all states accepting strings σ' whose Levenshtein distance $\Delta(\sigma, \sigma') \le d_{\max}$.

To avoid creating a parallel bundle of arcs for each insertion and substitution point, we instead decorate each arc with a nominal predicate, accepting or rejecting σ_i. To distinguish this nominal variant from the original construction, we highlight the modified rules in orange below.

$$\frac{i \in [0,n] \quad j \in [1,d_{\max}]}{(q_{i,j-1} \xrightarrow{[\neq \sigma_{i+1}]} q_{i,j}) \in \delta} \qquad \frac{i \in [1,n] \quad j \in [1,d_{\max}]}{(q_{i-1,j-1} \xrightarrow{[\neq \sigma_i]} q_{i,j}) \in \delta}$$

$$\frac{i \in [1,n] \quad j \in [0,d_{\max}]}{(q_{i-1,j} \xrightarrow{[= \sigma_i]} q_{i,j}) \in \delta} \qquad \frac{d \in [1,d_{\max}] \quad i \in [d+1,n] \quad j \in [d,d_{\max}]}{(q_{i-d-1,j-d} \xrightarrow{[= \sigma_i]} q_{i,j}) \in \delta}$$

Nominalizing the NFA eliminates the creation of $e = 2(|\Sigma| - 1) \cdot |\sigma| \cdot d_{\max}$ unnecessary arcs over the entire Levenshtein automaton and drastically reduces the representation size, but does not affect the underlying semantics. Thus, it is important to first nominalize the automaton before proceeding to avoid a large blowup in the intermediate grammar.

As a concrete example, suppose we have the string, $\sigma =$ ( )) and wish to balance the parentheses. We will initially have the Levenshtein automaton, A, depicted in Fig. 3. To check for non-emptiness, we will perform the following procedure. Suppose we have a CNF CFG, $G' = \{S \to LR, S \to LF, S \to SS, F \to SR, L \to (, R \to)\}$ and let us assume an ordering of S, F, L, R on V.

First, we need to order the automata states by increasing longest-path distance from q_0. One approach would be to topologically sort the adjacency matrix. While some form of sorting is unavoidable for arbitrary ANFAs, if we know ahead of time that our structure is a Levenshtein automaton, we can simply enumerate its state space by increasing Manhattan distance from the origin. So, a valid ordering on Q would be $q_{00}, q_{01}, q_{10}, q_{11}, q_{20}, q_{21}, q_{30}, q_{31}$. Now, we want to compute $L(G') \cap L(A)$.

Under such an ordering, the adjacency matrix takes an upper triangular form. This forms the template for the initial parse chart. The initial parse chart, M_0 (Fig. 6) and its fixpoint M_∞, post convergence, (Fig. 7) are shown. Each entry of this chart corresponds to a vector of expressions $E^{|V|}$ with at least one expression denoting a nonempty language. The adjacency and reachability matrices will always cover the expression vectors of the initial and final parse charts. In other words, we can always skip empty states in the reachability matrix.

Since $M_\infty[q_{00}, q_{31}, 0] = \blacksquare$, this implies that $L(A) \cap L(G') \neq \varnothing$, hence $\mathrm{LED}(\sigma, G) = 1$. Using the same matrix, we can then perform a second pass over the nonempty entries to construct regular expression vectors representing finite languages for each constituent and decode a witness using the Brzozowski derivative.

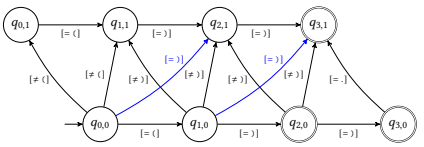


Fig. 3. Simple Levenshtein automaton.

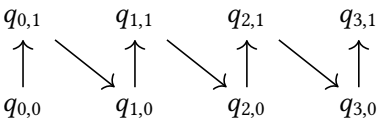


Fig. 4. Pairing function over $L(\sigma : \Sigma^3, 1)$.

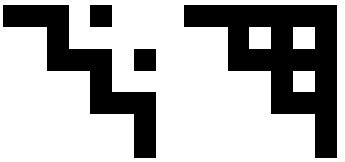


Fig. 5. Adjacency and reachability matrix.

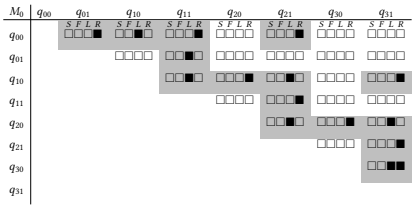


Fig. 6. Initial parse chart configuration.

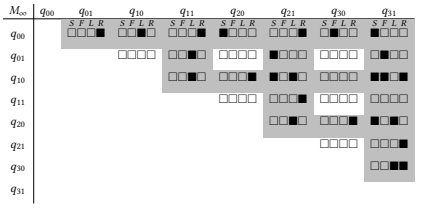


Fig. 7. Final parse chart configuration.

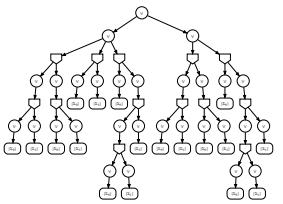


Fig. 8. Regular expression denoting $\mathcal{L}(G_\cap)$.

7 MEASURING THE LANGUAGE INTERSECTION

We will now attempt to put a probability distribution over the language intersection. We will start with a few cursory but illumative approaches, then proceed towards a more refined solution.

7.1 Exact enumeration

A brute force solution would be to generate every path and rank every one by its probability. It should be obvious why is unviable due its worst case complexity, but bears mentioning due to its global optimality. In certain cases, it can be realized when the intersection language is small.

To enumerate, we first need $|\mathcal{L}(e)|$, which is denoted $|e|$ for brevity.

Definition 7.1 (Cardinality). $$|e| : E \to \mathbb{N} = \begin{cases} 1 & \text{if } R \in \Sigma \\ x \times z & \text{if } e = x \cdot z \\ x + z & \text{if } e = x \vee z \end{cases}$$

Theorem 7.2 (Enumeration). *To enumerate, invoke* $\bigcup_{i=0}^{|R|} \{\mathit{enum}(R, i)\}$*:*

$$\mathit{enum}(e, n) : E \times \mathbb{N} \to \Sigma^* = \begin{cases} e & \text{if } R \in \Sigma \\ \mathit{enum}\left(x, \lfloor \frac{n}{|z|} \rfloor\right) \cdot \mathit{enum}\left(z,\ n \bmod |z|\right) & \text{if } e = x \cdot z \\ \mathit{enum}\left((x, z)_{\min(1, \lfloor \frac{n}{|x|} \rfloor)}, n - |x| \min(1, \lfloor \frac{n}{|x|} \rfloor)\right) & \text{if } e = x \vee z \end{cases}$$

7.2 Mode collapse

Ordinarily, we would use top-down PCFG sampling, however in the case of non-recursive CFGs, this method is highly degenerate, exhibiting poor sample diversity. Consider an illustrative pathological case for top-down ancestral (TDA) sampling:

$$S \to A\ B\ (0.9999) \qquad S \to C\ C\ (0.0001)$$

$$A \to a\ (1) \qquad B \to b\ (1) \qquad C \to a \left(\frac{1}{26}\right) \mid \ldots \mid z \left(\frac{1}{26}\right)$$

TDA sampling will almost always generate the string *ab*, but most of the language is concealed in the hidden branch, $S \to CC$. Although contrived example, it illustrates precisely why TDA sampling is unviable: we want a sampler that matches the true distribution over the finite CFL, not the PCFG's local approximation thereof.

7.3 Ambiguity

Another approach would be to sample trees and rerank them by their PCFG score. More pernicious is the issue of ambiguity. Since the CFG can be ambiguous, this causes certain repairs to be overrepresented, resulting in a subtle bias. Consider for example,

Lemma 7.3. *If the FSA, α, is ambiguous, then the intersection grammar, $G_\cap$, can be ambiguous.*

Proof. Let ℓ be the language defined by $G = \{S \to LR, L \to (, R \to)\}$, where $\alpha = L(\underset{\sim}{\sigma}, 2)$, the broken string $\underset{\sim}{\sigma}$ is)(, and $\mathcal{L}(G_\cap) = \ell \cap \mathcal{L}(\alpha)$. Then, $\mathcal{L}(G_\cap)$ contains the following two identical repairs:)() with the parse $S \to q_{00}Lq_{21}\ q_{21}Rq_{22}$, and () with the parse $S \to q_{00}Lq_{11}\ q_{11}Rq_{22}$. □

We would like the underlying sample space to be a proper set, *not* a multiset.

REFERENCES

[1] Janusz A Brzozowski. 1964. Derivatives of regular expressions. Journal of the ACM (JACM) 11, 4 (1964), 481–494.

[2] Noam Chomsky. 1959. On certain formal properties of grammars. Information and control 2, 2 (1959), 137–167.

[3] Joshua Goodman. 1999. Semiring parsing. Computational Linguistics 25, 4 (1999), 573–606. https://aclanthology.org/J99-4004.pdf

[4] Dick Grune and Ceriel J. H. Jacobs. 2008. Parsing as Intersection. Springer New York, New York, NY, 425–442. https://doi.org/10.1007/978-0-387-68954-8_13

[5] Arto Salomaa. 1973. Formal languages. Academic Press, New York. 59–61 pages.

[6] Klaus U Schulz and Stoyan Mihov. 2002. Fast string correction with Levenshtein automata. International Journal on Document Analysis and Recognition 5 (2002), 67–85.

A LEVENSHTEIN AUTOMATA MATRICES

These are useful for visually checking different implementations.

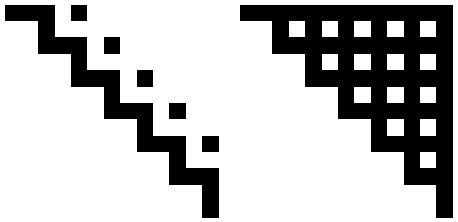


Fig. 9. Lev($|\sigma|$=6, Δ=1) adjacency and reachability matrices.

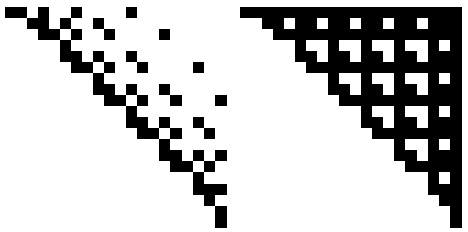


Fig. 10. Lev($|\sigma|$=6, Δ=2) adjacency and reachability matrices.

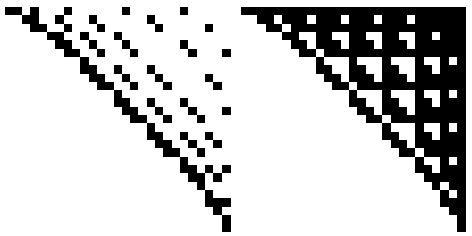


Fig. 11. Lev($|\sigma|$=6, Δ=3) adjacency and reachability matrices.

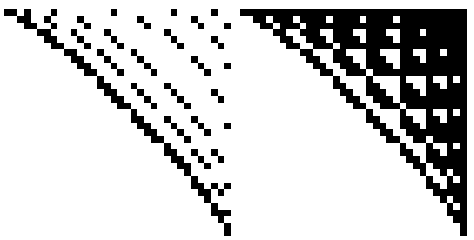


Fig. 12. Lev($|\sigma|$=6, Δ=4) adjacency and reachability matrices.

B LEVENSHTEIN AUTOMATA MINIMALITY

It is reasonable to ask whether the Levenshtein automaton defined is minimal, in the sense of whether there exists an automaton with fewer states than A yet still generates $\mathcal{L}(G_\cap)$ when intersected with $\mathcal{L}(G)$. In other words, given G and $\underset{\sim}{\sigma}$, is there an A' such that $|Q_{A'}| < |Q_A|$ yet $\mathcal{L}(G) \cap \mathcal{L}(A') = \mathcal{L}(G) \cap \mathcal{L}(A)$ still holds? In fact, there is a trivial example:

Theorem B.1. *Let $Q_{A'}$ be defined as $Q_A \setminus \{q_{n,0}\}$.*

Since $q_{n,0}$ accepts the original string $\underset{\sim}{\sigma}$ which by definition is not in $\mathcal{L}(G)$, we can immediately rule out this state. Moreover, we can define a family of automata with strictly fewer states than the full LBH construction by making the following observation: if we can prove one edit must occur before the last s tokens, we can rule out the last s states absorbing editless trajectories.

Theorem B.2. *$\varnothing = \mathcal{L}(\underset{\sim}{\sigma}_{1\ldots(n-s)} \cdot \Sigma^s) \cap \mathcal{L}(G)$ implies the states $[q_{n-i,0}]_{i\in 1\ldots s}$ are unnecessary.*

Likewise, if we expend our entire edit budget in the first p tokens, we will be unable to recover in a string where at least one repair must occur after the first p tokens.

Theorem B.3. *$\varnothing = \mathcal{L}(\Sigma^p \cdot \underset{\sim}{\sigma}_p) \cap \mathcal{L}(G)$ implies the states $[q_{i,d_{\max}}]_{i\in 0\ldots p}$ are unnecessary.*

Therefor, we can eliminate $p+s$ states from A by proving emptiness of $\mathcal{L}(\Sigma^p \cdot \underset{\sim}{\sigma}_{p\ldots(n-s)} \cdot \Sigma^s) \cap \mathcal{L}(G)$, without affecting $\mathcal{L}(G_\cap)$. Pictorially,

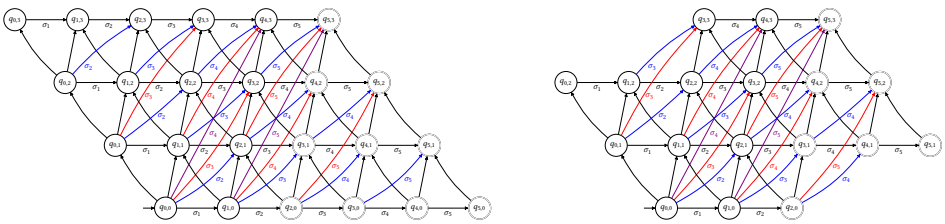


Fig. 13. Levenshtein NFA before and after pruning.

Pruned L-NFA for the broken string $\underset{\sim}{\sigma}$ = `[ ( + ) ]` with $G = \{S \to (S) \mid [S] \mid S + S \mid 1\}$.

`_ _ + ) ]` ❌ ∧ `_ _ _ ) ]` ✅

`[ ( + _ _` ❌ ∧ `[ ( _ _ _` ✅