# Syntax Repair as Idempotent Tensor Completion

Breandan Considine[1], Jin Guo[1], and Xujie Si[2]

[1] McGill University, Montréal, QC H2R 2Z4, Canada

[2] University of Toronto, Toronto, ON, M5S 1A1 Canada

**Abstract.** We introduce a new technique for correcting syntax errors in arbitrary context-free languages. To do so, we reduce CFL recognition onto a Boolean tensor completion and compare various techniques for introducing the holes, and solving for their inhabitants. Our technique has practical applications for real-time syntax correction in programming languages.

**Keywords:** Error correction · CFL reachability · Langauge games.

## 1 Introduction

Syntax repair is the problem of taking a grammar and a malformed string, and modifying the string so it conforms to the grammar. Prior work has been devoted to fixing syntax errors using handcrafted heuristics. We take a first-principles approach that makes no assumptions about the string or grammar and focuses on accuracy and end-to-end latency. The result is a tool that is applicable to any context-free and conjunctive language, and which is provably sound and complete up to a Levenshtein bound.

### 1.1 Problem

Syntax repair can be treated as a language intersection problem between a context-free language (CFL) and a regular language.

**Definition 1 (Bounded Levenshtein-CFL reachability).** *Given a CFL $\ell$ and an invalid string $\underset{\sim}{\sigma} : \ell^{\complement}$, the BCFLR problem is to find every valid string reachable within $d$ edits of $\underset{\sim}{\sigma}$, i.e., letting $\Delta$ be the Levenshtein metric and $L(\underset{\sim}{\sigma}, d) \coloneqq \{\sigma \mid \Delta(\underset{\sim}{\sigma}, \sigma) \leq d\}$, we seek to find $L(\underset{\sim}{\sigma}, d) \cap \ell$.*

To solve this problem, we will first pose a simpler problem that only considers intersections with a finite language, then turn our attention back to BCFLR.

**Definition 2 (Porous completion).** *Let $\underline{\Sigma} \coloneqq \Sigma \cup \{\_\}$, 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 $\{\sigma' \mid \sigma' \sqsubseteq \sigma\}$ as $H(\sigma)$. Given a porous string, $\sigma : \underline{\Sigma}^*$ we seek all syntactically admissible inhabitants, i.e., $A(\sigma) \coloneqq H(\sigma) \cap \ell$.*

$A(\sigma)$ is often a large-cardinality set, so we want a procedure which returns the most likely members first, without exhaustive enumeration. More precisely,

**Definition 3 (Ranked repair).** *Given a finite language $\ell^\cap := L(\underset{\sim}{\sigma}, d) \cap \ell$ and a probabilistic language model $P_\theta : \Sigma^* \to [0,1] \subset \mathbb{R}$, the ranked repair problem is to find the top-k repairs by likelihood under the language model. That is,*

$$R(\ell^\cap, P_\theta) := \underset{\{\boldsymbol{\sigma} | \boldsymbol{\sigma} \subseteq \ell^\cap, |\boldsymbol{\sigma}| \leq k\}}{\operatorname{argmax}} \sum_{\sigma \in \boldsymbol{\sigma}} \prod_{i=1}^{|\sigma|} P_\theta(\sigma_i \mid \sigma_{1\ldots i})^{\frac{1}{|\sigma|}} \tag{1}$$

*We want a procedure $\hat{R}$, minimizing $\mathbb{E}_{G,\sigma}\big[D_{KL}(\hat{R} \parallel R)\big]$ and wallclock runtime.*

Our key innovation and the core problem this paper tackles is, given $\underset{\sim}{\sigma}, d, P_\theta$, to approximate $R(\ell^\cap, P_\theta)$ while minimizing latency and maximizing accuracy. We will first give an example, then dive into the theory.

### 1.2 Background

Recall that a CFG is a quadruple consisting of terminals ($\Sigma$), nonterminals ($V$), productions ($P \colon V \to (V \mid \Sigma)^*$), and a start symbol, ($S$). Every CFG is reducible to *Chomsky Normal Form*, $P' \colon V \to (V^2 \mid \Sigma)$, in which every $P$ takes one of two forms, either $w \to xz$, or $w \to t$, where $w, x, z : V$ and $t : \Sigma$. For example:

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

Given a CFG, $G' : \mathbb{G} = \langle \Sigma, V, P, S \rangle$ in CNF, we can construct a recognizer $R : \mathbb{G} \to \Sigma^n \to \mathbb{B}$ for strings $\sigma : \Sigma^n$ as follows. Let $2^V$ be our domain, 0 be $\varnothing$, $\oplus$ be $\cup$, and $\otimes$ be defined as:

$$X \otimes Z := \{\ w \mid \langle x, z \rangle \in X \times Z, (w \to xz) \in P\ \} \tag{2}$$

If we define $\sigma_r := \{w \mid (w \to \sigma_r) \in P\}$, then construct a matrix with nonterminals on the superdiagonal representing each token, $M_{r+1=c}(G', e) := \sigma_r$ and solve for the fixpoint $M_{i+1} = M_i + M_i^2$,

$$M_0 := \begin{pmatrix} \varnothing & \sigma_1 & \varnothing & \varnothing \\ & \ddots & \ddots & \\ & & \ddots & \varnothing \\ & & & \sigma_n \\ \varnothing & \cdots & & \varnothing \end{pmatrix} \Rightarrow \begin{pmatrix} \varnothing & \sigma_1 & \Lambda & \varnothing \\ & \ddots & \ddots & \\ & & \ddots & \Lambda \\ & & & \sigma_n \\ \varnothing & \cdots & & \varnothing \end{pmatrix} \Rightarrow \ldots \Rightarrow M_\infty = \begin{pmatrix} \varnothing & \sigma_1 & \Lambda & \Lambda_\sigma^* \\ & \ddots & \ddots & \\ & & \ddots & \Lambda \\ & & & \sigma_n \\ \varnothing & \cdots & & \varnothing \end{pmatrix}$$

we obtain the recognizer, $R(G', \sigma) := [S \in \Lambda_\sigma^*] \Leftrightarrow [\sigma \in \mathcal{L}(G)]$ [3].

Since $\bigoplus_{c=1}^{n} M_{r,c} \otimes M_{c,r}$ has cardinality bounded by $|V|$, it can be represented as $\mathbb{Z}_2^{|V|}$ using the characteristic function, $\mathbb{1}$. A concrete example is shown in § 1.3.

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

### 1.3 Example

Let us consider an example with two holes, $\sigma = 1 \_ \_$, and the grammar being $G \coloneqq \{S \to NON, O \to + \mid \times, N \to 0 \mid 1\}$. This can be rewritten into CNF as $G' \coloneqq \{S \to NL, N \to 0 \mid 1, O \to \times \mid +, L \to ON\}$. Using the algebra where $\oplus \coloneqq \cup$, $X \otimes Z \coloneqq \left\{\ w \mid \langle x, z\rangle \in X \times Z, (w \to xz) \in P\ \right\}$, the fixpoint $M' = M + M^2$ can be computed as follows, shown in the leftmost column:

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

The same procedure can be translated, without loss of generality, into the bit domain ($\mathbb{Z}_2^{|V|}$) using a lexicographic ordering, however these both are recognizers. That is to say, $[S \in V_{0,3}] \Leftrightarrow [V_{0,3,3} = 1] \Leftrightarrow [A(\sigma) \neq \varnothing]$. Since $V_{0,3} = \{S\}$, we know there is at least one $\sigma' \in A(\sigma)$, but $M_\infty$ does not reveal its identity.

In order 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 $X \boxtimes Z \coloneqq [X_2 \wedge Z_1, \bot, \bot, X_1 \wedge Z_0]$ and $X \boxplus Z \coloneqq [X_i \vee Z_i]_{i \in [0,|V|)}$. Since the unit nonterminals $O, N$ can only occur on the superdiagonal, they may be safely ignored by $\otimes$. To solve for $M_\infty$, we proceed by first computing $V_{0,2}, V_{0,3}$ as follows:

$$
\begin{aligned}
V_{0,2} &= V_{0,j} \cdot V_{j,2} = V_{0,1} \boxtimes V_{1,2} && (3)\\
&= [L \in V_{0,2}, \bot, \bot, S \in V_{0,2}] && (4)\\
&= [O \in V_{0,1} \wedge N \in V_{1,2}, \bot, \bot, N \in V_{0,1} \wedge L \in V_{1,2}] && (5)\\
&= [V_{0,1,2} \wedge V_{1,2,1}, \bot, \bot, V_{0,1,1} \wedge V_{1,2,0}] && (6)
\end{aligned}
$$

$$
\begin{aligned}
V_{1,3} &= V_{1,j} \cdot V_{j,3} = V_{1,2} \boxtimes V_{2,3} && (7)\\
&= [L \in V_{1,3}, \bot, \bot, S \in V_{1,3}] && (8)\\
&= [O \in V_{1,2} \wedge N \in V_{2,3}, \bot, \bot, N \in V_{1,2} \wedge L \in V_{2,3}] && (9)\\
&= [V_{1,2,2} \wedge V_{2,3,1}, \bot, \bot, V_{1,2,1} \wedge V_{2,3,0}] && (10)
\end{aligned}
$$

Now we can solve for $V_{0,3}$ by taking the bitwise dot product:

$$V_{0,3} = V_{0,j} \cdot V_{j,3} = V_{0,1} \boxtimes V_{1,3} \boxplus V_{0,2} \boxtimes V_{2,3} \tag{11}$$
$$= [V_{0,1,2} \wedge V_{1,3,1} \vee V_{0,2,2} \wedge V_{2,3,1}, \bot, \bot, V_{0,1,1} \wedge V_{1,3,0} \vee V_{0,2,1} \wedge V_{2,3,0}] \tag{12}$$

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

$$V_{0,3,3} = V_{0,1,1} \wedge V_{1,3,0} \vee V_{0,2,1} \wedge V_{2,3,0} \tag{13}$$
$$= V_{0,1,1} \wedge (V_{1,2,2} \wedge V_{2,3,1}) \vee V_{0,2,1} \wedge \bot \tag{14}$$
$$= V_{0,1,1} \wedge V_{1,2,2} \wedge V_{2,3,1} \tag{15}$$
$$= [N \in V_{0,1}] \wedge [O \in V_{1,2}] \wedge [N \in V_{2,3}] \tag{16}$$

Now we know that $\sigma = 1\ \underline{\mathrm{O}}\ \underline{\mathrm{N}}$ is a valid solution, and therefor we can take the product $\{1\} \times \sigma_r^{-1}(O) \times \sigma_r^{-1}(N)$ to recover the admissible set, yielding $A(\sigma) = \{1+0, 1+1, 1\times 0, 1 \times 1\}$. In this case, since $G$ is unambiguous, there is only one parse tree satisfying $V_{0,|\sigma|,|\sigma|}$, but in general, there can be multiple valid parse trees, in which case we can decode them incrementally.

### 1.4 Semiring Algebras

There are a number of strategies to tackling this problem. A first approach requires solving for $A(\sigma)$ using a semiring algebra, and propagating the values from the bottom-up as a string to a list of strings. Letting $D = V \to \mathcal{P}(\Sigma^*)$, we define $\oplus, \otimes : D \times D \to D$. Initially, we have $p(s : \Sigma) \coloneqq \{v \mid [v \to s] \in P\}$ and $p(\_) := \bigcup_{s \in \Sigma} p(s)$, then we compute the fixpoint using the following algebra:

$$X \oplus Z := \{v \to \big(X(v) \cup Z(v)\big) \mid v \in V\} \tag{17}$$

$$X \otimes Z := \bigoplus_{w,x,z} \big\{w \to (l + r) \mid [w \to xz] \in P, \langle l, r\rangle \in X(x) \times Z(x)\big\} \tag{18}$$

After the fixpoint $M_\infty$ is attained, the solutions can be read off via $M_\infty[0, |\sigma|](S)$. The issue here is an exponential growth in cardinality when eagerly computing the Cartesian product, which becomes impractical for even small strings. We can make this encoding more compact by propagating an algebraic data type (ADT) $\mathbb{T}_2$ using the operations $\oplus, \otimes : 2^{\mathbb{T}_2} \times 2^{\mathbb{T}_2} \to 2^{\mathbb{T}_2}$ as follows:

$$X \oplus Z := \{\mathbb{T}_2(k, Q_x \cup Q_z) \mid (k, Q)_X \bowtie_k (k, Q)_Z\} \tag{19}$$

$$X \otimes Z := \bigoplus_{w,x,z} \big\{\mathbb{T}_2(w, \{\langle T_x, T_z\rangle\}) \mid [w \to xz] \in P, x \in \pi_1(X), z \in \pi_1(Z)\big\} \tag{20}$$

Decoding, then becomes a matter of enumerating binary trees from the ADT using a recursive choice function that emits a sequence of strings satisfying $A(\sigma)$, with the type signature $\mathcal{C} : \mathbb{T}_2 \to (\mathbb{N} \to \Sigma^*)$ defined as follows:

$$\mathcal{C}(t : \mathbb{T}_2) := \begin{cases} \pi_1(t) & \text{if } \pi_2(t) = \varnothing, \text{ or} \\ \{x + z \mid \langle X, Z\rangle \in \pi_2(t), x \in \mathcal{C}(X), z \in \mathcal{C}(Z)\} & \text{otherwise.} \end{cases}$$

### 1.5 Bounded CFL Reachability

Now, let us return to the problem of bounded CFL reachability. A well-known result in FL theory is that the class of context-free languages are closed under intersection with regular languages, i.e.,

$$\ell_1 : \textsc{Reg}, \ell_2 : \textsc{Cfl} \vdash \text{ there exists } G \text{ s.t. } L(G) : \textsc{Cfl} \text{ and } L(G) = \ell_1 \cap \ell_2 \quad (21)$$

To compute the intersection between a CFG and a regular grammar, there is a standard construction, which can be attributed to Beigel and Gasarch: …

### 1.6 Ranking

Since the number of solutions can be very large, we can use a language model to rank the results maximizing likelihood, or minimizing perplexity, subject to the constraints. This ranking can be used to guide the propagation, sample the choice function, sample hole locations or as a post-processing step after all solutions have been found.

## References

1. Author, B.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, . Last accessed 4 Oct 2017