On Iterated Scattered Deletion

Michael Domaratzki*
School of Computing Queen's University,
Kingston, ON K7L 3N6 Canada
email: domaratz@cs.queensu.ca

In this note, we solve an open problem of Ito *et al.* [2] on iterated scattered deletion.

Note: Since this note has appeared in Bull. EATCS, I have been informed that this problem has been previously solved; see Ito and Silva [1], where the authors show that there exists a regular language R such that $(\rightsquigarrow)^+(R)$ is not a CFL. I am grateful to Masami Ito for pointing me to this reference.

Let Σ be an alphabet. The scattered deletion [3, 5] of two words $x, y \in \Sigma^*$, denoted by $x \rightsquigarrow y$, is defined as

$$x \rightsquigarrow y = \{x_1 x_2 \cdots x_k \ : \ y = y_1 \cdots y_{k-1}; x = x_1 y_1 x_2 y_2 \cdots x_{k-1} y_{k-1} x_k; x_i, y_i \in \Sigma^*\}.$$

Thus, $x \rightsquigarrow y$ is the set of words which result from deleting y as a scattered subword from x. We extend this operation to languages $L_1, L_2 \subseteq \Sigma^*$ as follows:

$$L_1 \rightsquigarrow L_2 = \bigcup_{x \in L_1} \bigcup_{y \in L_2} x \rightsquigarrow y.$$

For unexplained notions in formal language and automata theory, please see Yu [6]. For languages $L_1, \cdots, L_k$, we use the notation $\prod_{i=1}^k L_i = L_1 L_2 \cdots L_k$.

We now define an iterated scattered deletion operation [2]. Let $i \geq 1$, $L \subseteq \Sigma^*$. Then $(\rightsquigarrow)^i(L)$ is defined recursively as follows:

$$\begin{aligned} (\rightsquigarrow)^1(L) &= L; \\ (\rightsquigarrow)^{i+1}(L) &= (\rightsquigarrow)^i(L) \rightsquigarrow ((\rightsquigarrow)^i(L) \cup \{\epsilon\}) \quad \forall i \geq 1. \end{aligned}$$

Then the iterated scattered deletion operator $(\rightsquigarrow)^+(L)$ is given by

$$(\rightsquigarrow)^+(L) = \bigcup_{i \geq 1} (\rightsquigarrow)^i(L).$$

*Research supported in part by an NSERC PGS-B graduate scholarship.

We also define an auxiliary operation $L_1[\leadsto]^i L_2$ which is defined recursively for all $i \geq 0$ as follows:

$$\begin{aligned} L_1[\leadsto]^0 L_2 &= L_1; \\ L_2[\leadsto]^{i+1} L_2 &= (L_1[\leadsto]^i L_2) \leadsto L_2 \quad \forall i \geq 1. \end{aligned}$$

We then set

$$L_1[\leadsto]^* L_2 = \bigcup_{i \geq 0} L_1[\leadsto]^i L_2.$$

Ito *et al.* [2] asked whether the regular languages are closed under $(\leadsto)^+$. We show that they are not.

Let $k \geq 2$ be arbitrary, and let $\Sigma_k = \{\alpha_i, \beta_i, \gamma_i, \eta_i\}_{i=1}^k$. Then we define $L_k \subseteq \Sigma_k^*$ as

$$L_k = \prod_{i=1}^k (\alpha_i \beta_i)^* \prod_{i=1}^k (\gamma_i \eta_i)^* + \bigcup_{i=1}^k \beta_i \eta_i.$$

We claim that

$$(\leadsto)^+(L_k) \cap \prod_{i=1}^k \alpha_i^+ \prod_{i=1}^k \gamma_i^+ = \{\alpha_1^{i_1} \alpha_2^{i_2} \cdots \alpha_k^{i_k} \gamma_1^{i_1} \gamma_2^{i_2} \cdots \gamma_k^{i_k} \ : \ i_j \geq 1\}. \quad (1)$$

and that $(\leadsto)^+(L_k)$ cannot be expressed as the intersection of $k-1$ context-free languages.

We first establish (1). Let $(i_1, i_2, \cdots, i_k) \in \mathbb{N}^k$. Then note that

$$\prod_{j=1}^k \alpha_j^{i_j} \prod_{j=1}^k \gamma_j^{i_j} \in (\cdots (\prod_{j=1}^k (\alpha_j \beta_j)^{i_j} \prod_{j=1}^k (\gamma_j \eta_j)^{i_j})[\leadsto]^{i_1} \beta_1 \eta_1) \cdots)[\leadsto]^{i_k} \beta_k \eta_k.$$

This establishes the right-to-left inclusion of (1). We now show the reverse inclusion. First, note that if $\alpha \in (\leadsto)^+(L_k)$, then we can write $\alpha = x_1 x_2 \cdots x_k y_1 y_2 \cdots y_k$ where $x_i \in \{\alpha_i, \beta_i\}^*$ and $y_i \in \{\gamma_i, \eta_i\}^*$. To prove the left-to-right inclusion of (1), we will require the following stronger claim:

Claim 1 *Let $x_1 x_2 \cdots x_k y_1 y_2 \cdots y_k \in (\leadsto)^+(L_k)$ where $x_i \in \{\alpha_i, \beta_i\}^*$ and $y_i \in \{\gamma_i, \eta_i\}^*$ for all $1 \leq i \leq k$. Then for all $1 \leq i \leq k$, the following equalities hold:*

$$|x_i|_{\alpha_i} - |x_i|_{\beta_i} = |y_i|_{\gamma_i} - |y_i|_{\eta_i}. \quad (2)$$

Proof. Let $z = x_1 x_2 \cdots x_k y_1 y_2 \cdots y_k \in (\leadsto)^+(L_k)$. Then there exists some $i \geq 1$ such that $z \in (\leadsto)^i(L_k)$. The proof is by induction on i. For $i = 1$, $z \in L_k$. Thus, we see that either

(a) for all $1 \leq \ell \leq k$, $x_\ell = (\alpha_\ell \beta_\ell)^{j_\ell}$ for some $j_\ell \geq 0$ and $y_\ell = (\gamma_\ell \eta_\ell)^{j'_\ell}$ for some $j'_\ell \geq 0$, in which case $|x_\ell|_{\alpha_\ell} - |x_\ell|_{\beta_\ell} = 0 = |y_\ell|_{\gamma_\ell} - |y_\ell|_{\eta_\ell}$; or

(b) $z = \beta_\ell \eta_\ell$ for some $1 \leq \ell \leq k$. Thus, $|x_\ell|_{\alpha_\ell} - |x_\ell|_{\beta_\ell} = -1 = |y_\ell|_{\gamma_\ell} - |y_\ell|_{\eta_\ell}$ and $x_j = y_j = \epsilon$ for all $j \neq \ell$ with $1 \leq j \leq k$.

Thus, the result holds for $i = 1$.

Assume the claim holds for all natural numbers less than i. Let $z \in (\rightsquigarrow)^i(L_k)$. Then there exists some $\theta \in (\rightsquigarrow)^{i-1}(L_k)$ and $\zeta \in (\rightsquigarrow)^{i-1}(L_k) + \epsilon$ such that $z \in \theta \rightsquigarrow \zeta$. If $\zeta = \epsilon$, then $z = \theta$ and the result holds by induction. Thus, let

$$\begin{aligned} \theta &= u_1 u_2 \cdots u_k v_1 v_2 \cdots v_k \\ \zeta &= s_1 s_2 \cdots s_k t_1 t_2 \cdots t_k \end{aligned}$$

so that $u_\ell, s_\ell \in \{\alpha_\ell, \beta_\ell\}^*$ and $v_\ell, t_\ell \in \{\gamma_\ell, \eta_\ell\}^*$ for all $1 \leq \ell \leq k$. Then note that for all $1 \leq \ell \leq k$,

$$\begin{aligned} |x_\ell|_{\alpha_\ell} &= |u_\ell|_{\alpha_\ell} - |s_\ell|_{\alpha_\ell}; \\ |x_\ell|_{\beta_\ell} &= |u_\ell|_{\beta_\ell} - |s_\ell|_{\beta_\ell}; \\ |y_\ell|_{\gamma_\ell} &= |v_\ell|_{\gamma_\ell} - |t_\ell|_{\gamma_\ell}; \\ |y_\ell|_{\eta_\ell} &= |v_\ell|_{\eta_\ell} - |t_\ell|_{\eta_\ell}. \end{aligned}$$

Thus, by induction, we can easily establish that the desired equalities hold. ■

We now show that $(\rightsquigarrow)^+(L_k)$ cannot be expressed as the intersection of $k-1$ context-free languages. Let CFL_k be the class of languages which are expressible as the intersection of k CFLs. The following lemma is obvious, since the CFLs are closed under intersection with regular languages.

Lemma 2 *CFL_k is closed under intersection with regular languages.*

We will also require the following lemma:

Lemma 3 *Let $L_1, L_2 \in CFL_k$ be such that there exist disjoint regular languages R_1, R_2 such that $L_i \subseteq R_i$ for $i = 1, 2$. Then $L_1 \cup L_2 \in CFL_k$.*

Proof. Let $L_1 = X_1 \cap \cdots \cap X_k$ and $L_2 = Y_1 \cap \cdots \cap Y_k$. Then without loss of generality, we may assume that $X_j \subseteq R_1$ and $Y_j \subseteq R_2$ for $1 \leq j \leq k$; if not, we may replace X_i with $X_i \cap R_1$ and Y_i with $Y_i \cap R_2$ as necessary. Both intersections are still CFLs.

Thus, note that $L_1 \cup L_2 = (X_1 \cup Y_1) \cap \cdots \cap (X_k \cup Y_k)$. As $X_i \cup Y_i \in CFL$, the result immediately follows. ■

The following result is due to Liu and Weiner [4, Thm. 8]:

Theorem 4 *Let $k \geq 2$. Let $L''_k = \{\alpha_1^{i_1} \alpha_2^{i_2} \cdots \alpha_k^{i_k} \alpha_1^{i_1} \alpha_2^{i_2} \cdots \alpha_k^{i_k} \ : \ i_j \geq 0\}$. Then $L'_k \in CFL_k - CFL_{k-1}$.*

However, we prove the following corollary, which will be more useful to us:

Corollary 5 *Let* $L'_k = \{\alpha_1^{i_1}\alpha_2^{i_2}\cdots\alpha_k^{i_k}\alpha_1^{i_1}\alpha_2^{i_2}\cdots\alpha_k^{i_k} \;:\; i_j \geq 1\}$. *Then* $L'_k \in CFL_k - CFL_{k-1}$.

Proof. The sufficiency of k intersections is obvious, by Lemma 2. We prove only the necessity of k intersections. The proof is by induction. For $k = 2$, the result can be established by the pumping lemma. Let $S \subset [k]$. Denote by $L_k^{(S)}$ the language

$$L_k^{(S)} = \{\prod_{j\in S}\alpha_j^{i_j}\prod_{j\in S}\alpha_j^{i_j} \;:\; i_j \geq 1\}.$$

Further, note that

$$L_k^{(S)} \subseteq (\prod_{j\in S}\alpha_j^{+})^2.$$

Let $R_S = (\prod_{j\in S}\alpha_j^{+})^2$. Then note that $R_S \cap R_{S'} = \emptyset$ for $S, S' \subseteq [k]$ (including the possibility that $S = [k]$) with $S \neq S'$.

By induction, if $S \subset [k]$, where the inclusion is proper, then $L_k^{(S)} \in CFL_{k-1}$.

Assume that L'_k can be expressed as the intersection of $k-1$ CFLs. We then note that

$$L''_k = L'_k \cup \bigcup_{S\subsetneq[k]} L_k^{(S)}.$$

By Lemma 3, $L''_k \in CFL_{k-1}$, a contradiction. This completes the proof. ∎

Thus, we may state our main result:

Theorem 6 *For all* $k \geq 2$, *there exists an* $O(n^{2k-1})$*-density bounded regular language* L_k *such that* $(\rightsquigarrow)^+(L_k)$ *cannot be expressed as the intersection of* $k-1$ *context-free languages.*

Proof. Let $\Delta_k = \{\gamma_i, \alpha_i\}_{i=1}^k$. Let $h_k : \Delta_k^* \to \Delta_k^*$ be given by $h_k(\alpha_i) = h_k(\gamma_i) = \alpha_i$ for all $1 \leq i \leq k$. Let $D_k = (\rightsquigarrow)^+(L_k)$. Let $R_k = \prod_{i=1}^k \alpha_i^+ \prod_{i=1}^k \gamma_i^+$. If $D_k \in CFL_{k-1}$, then $D_k \cap R_k \in CFL_{k-1}$ as well, by Lemma 2. We now claim that this implies that $h_k(D_k \cap R_k)$ is in CFL_{k-1}.

Let $X_1, X_2, \ldots, X_{k-1} \in CFL$ be chosen so that

$$D_k \cap R_k = \cap_{i=1}^{k-1} X_i.$$

The inclusion $h_k(D_k \cap R_k) \subseteq \cap_{i=1}^{k-1} h_k(X_i)$ is easily verified. We now show the reverse inclusion. First, we may assume without loss of generality that $X_j \subseteq R_k$ for all $1 \leq j \leq k-1$. If not, let $X'_j = X_j \cap R_k$. By the closure properties of the CFLs, $X'_j \in CFL$, and we still have $D_k \cap R_k = \cap_{i=1}^{k-1} X'_i$.

Let $x \in \cap_{i=1}^{k-1} h_k(X_i)$. Let $y_i \in X_i$ be such that $h_k(y_i) = x$ for $1 \leq i \leq k-1$. By assumption, we can write

$$y_j = \prod_{i=1}^k \alpha_i^{\ell_i^{(j)}} \prod_{i=1}^k \gamma_i^{m_i^{(j)}}$$

for some $\ell_i^{(j)}, m_i^{(j)} \geq 1$ for $1 \leq i \leq k$ and $1 \leq j \leq k-1$. Thus, by definition of h_k,

$$h_k(y_j) = \prod_{i=1}^{k} \alpha_i^{\ell_i^{(j)}} \prod_{i=1}^{k} \alpha_i^{m_i^{(j)}},$$

for all $1 \leq j \leq k-1$. As $h_k(y_j) = x$, for all $1 \leq j \leq k-1$, we must have that $\ell_i^{(j)} = \ell_i^{(j')}$ and $m_i^{(j)} = m_i^{(j')}$ for $1 \leq i \leq k$ and all $1 \leq j, j' \leq k-1$. Thus $y_1 = \cdots = y_{k-1} \in \cap_{i=1}^{k-1} X_i$ and thus $x \in h_k(\cap_{i=1}^{k-1} X_i) = h_k(D_k \cap R_k)$. Thus,

$$h_k(D_k \cap R_k) = \cap_{i=1}^{k-1} h_k(X_i).$$

As $h_k(X_i)$ is a CFL for all $1 \leq i \leq k-1$, $h_k(D_k \cap R_k) \in CFL_{k-1}$. But now note that

$$h_k(D_k \cap R_k) = L_k'.$$

This contradicts Corollary 5. Thus, D_k cannot be expressed as the intersection of $k-1$ CFLs. ■

References

[1] Ito, M., and Silva, P. Remarks on Deletions, Scattered Deletions and Related Operations on Languages. In J. Howie *et al.*, eds., Semigroups and Applications (1998), 97–105.

[2] Ito, M., Kari, L., and Thierrin, G. Shuffle and scattered deletion closure of languages. *Theor. Comp. Sci. 245* (2000), 115–133.

[3] Kari, L. Deletion operations: Closure properties. *International Journal of Computer Mathematics 52* (1994), 23–42.

[4] Liu, L., and Weiner, P. An infinite hierarchy of intersections of context-free languages. *Math. Sys. Th. 7*, 2 (1973), 185–192.

[5] Sântean, L. Six arithmetic-like operation on languages. *Revue Roumaine de Linguistique – Cahiers de linguistique théoretique et applique 25* (1988), 65–73.

[6] Yu, S. Regular languages. In *Handbook of Formal Languages, Vol. I*, G. Rozenberg and A. Salomaa, Eds. Springer-Verlag, 1997, pp. 41–110.