Site-Directed Insertion: Decision Problems, Maximality and Minimality

DCFS 2018

Taylor J. Smith

Joint work with D.-J. Cho, Y.-S. Han, and K. Salomaa

School of Computing
Queen's University
Kingston, Ontario, Canada

July 27, 2018

Table of Contents

Table of Contents

Background

- The study of formal language operations has great applicability to the field of biocomputing.
  - e.g., Insertion/deletion/mutation in DNA strings.
- **Contextual insertion** is a variant of traditional insertion that considers **contexts** around words.
- Given a word y, a set of contexts C, and context words $u, v \in C$, we insert y into subwords of the form uv.
- We can refine contextual insertion by considering **overlapping contexts** between the inserted word and the context words.

Table of Contents

Definition

Given words x and y, the **site-directed insertion** of y into x is

$$x \overset{\text{sdi}}{\leftarrow} y = \{x_1 uzv x_2 \mid x = x_1 uv x_2, y = uzv, u \neq \epsilon, v \neq \epsilon\}.$$

- We now have two preconditions:
  - uv is a subword of x (same as contextual insertion)
  - u and v are nontrivial outfixes of y (new!)
- This operation was called "outfix-guided insertion" by Cho et al. in an earlier paper.

Definition

Given words x and y, the **maximal site-directed insertion** of y into x is

$$x \overset{\text{max-sdi}}{\longleftarrow} y = \{x_1 uzv x_2 \mid x = x_1 uv x_2, y = uzv, u \neq \epsilon, v \neq \epsilon\},$$

where there exist no nonempty suffix x_1' of x_1 and no nonempty prefix x_2' of x_2 such that $y = x_1' uz'v x_2'$ for some $z' \in \Sigma^*$.

Definition

Given words x and y, the **maximal site-directed insertion** of y into x is

$$x \overset{\text{max-sdi}}{\longleftarrow} y = \{x_1 uzv x_2 \mid x = x_1 uv x_2, y = uzv, u \neq \epsilon, v \neq \epsilon\},$$

where there exist no nonempty suffix x_1' of x_1 and no nonempty prefix x_2' of x_2 such that $y = x_1' uz'v x_2'$ for some $z' \in \Sigma^*$.

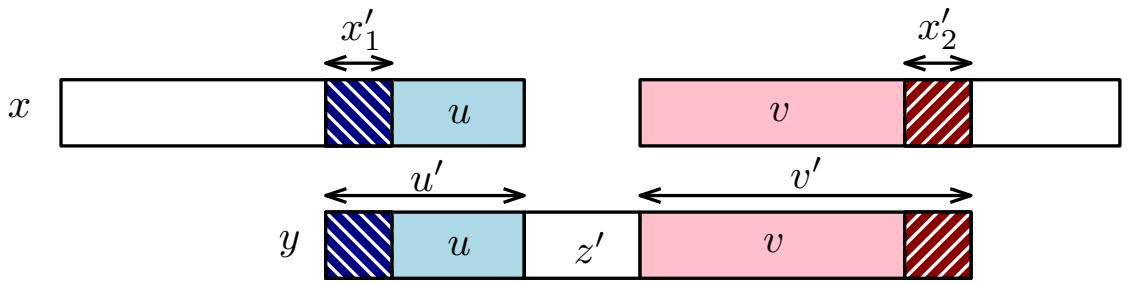


Figure: Non-maximal SDI

Definition

Given words x and y, the **maximal site-directed insertion** of y into x is

$$x \overset{\text{max-sdi}}{\longleftarrow} y = \{x_1 uzv x_2 \mid x = x_1 uv x_2, y = uzv, u \neq \epsilon, v \neq \epsilon\},$$

where there exist no nonempty suffix x_1' of x_1 and no nonempty prefix x_2' of x_2 such that $y = x_1' uz'v x_2'$ for some $z' \in \Sigma^*$.

Figure: Maximal SDI

Definition

Given words x and y, the **minimal site-directed insertion** of y into x is

$$x \overset{\text{min-sdi}}{\longleftarrow} y = \{x_1 uzv x_2 \mid x = x_1 uv x_2, y = uzv, u \neq \epsilon, v \neq \epsilon\},$$

where no proper nonempty suffix of u is a prefix of u,
and no proper nonempty prefix of v is a suffix of v.

Definition

Given words x and y, the **minimal site-directed insertion** of y into x is

$$x \overset{\text{min-sdi}}{\longleftarrow} y = \{x_1 uzv x_2 \mid x = x_1 uv x_2, y = uzv, u \neq \epsilon, v \neq \epsilon\},$$

where no proper nonempty suffix of u is a prefix of u, and no proper nonempty prefix of v is a suffix of v.

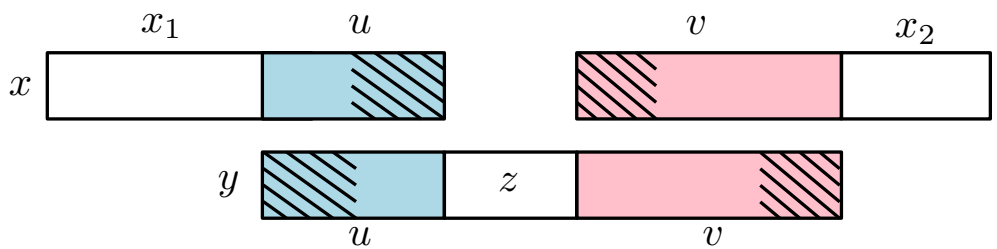


Figure: Non-minimal SDI

Definition

Given words x and y, the **minimal site-directed insertion** of y into x is

$$x \overset{\text{min-sdi}}{\longleftarrow} y = \{x_1 uzv x_2 \mid x = x_1 uv x_2, y = uzv, u \neq \epsilon, v \neq \epsilon\},$$

where no proper nonempty suffix of u is a prefix of u,
and no proper nonempty prefix of v is a suffix of v.

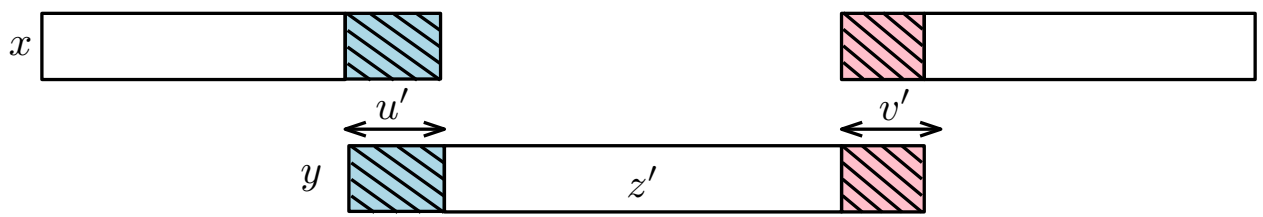


Figure: Minimal SDI

Table of Contents

- Although SDI is closed under regular languages, this is not the case for either max-SDI or min-SDI.

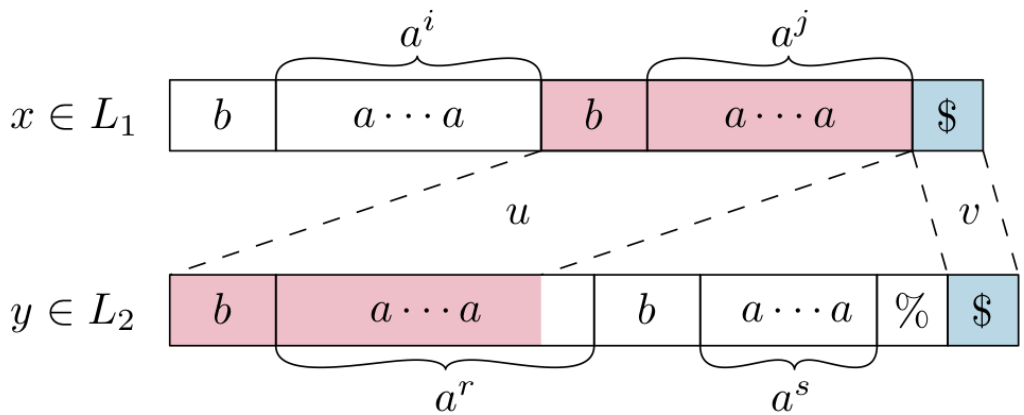
Theorem

Maximal/minimal site-directed insertion do not preserve regularity.

Proof Sketch

Inspired by the max/min chop constructions of Holzer et al.
Consider maximal SDI. Take $L_1 = ba^+ba^+\$$ and $L_2 = ba^+ba^+\%\$$.

$$(L_1 \overset{\text{max-sdi}}{\longleftarrow} L_2) \cap (ba^+)^3\%\$ = \{ba^m ba^n ba^k\%\$ \mid m \neq n \text{ or } k < n\}$$



- However, if we take a regular language and a finite language, then both max-SDI and min-SDI produce a regular language.

Theorem

Given a regular language R and a finite language L, the languages $R \overset{\text{max-sdi}}{\longleftarrow} L$, $L \overset{\text{max-sdi}}{\longleftarrow} R$, $R \overset{\text{min-sdi}}{\longleftarrow} L$, and $L \overset{\text{min-sdi}}{\longleftarrow} R$ are all regular.

Proof Sketch

Consider $R \overset{\text{max-sdi}}{\longleftarrow} L$. Let $\mathcal{A}$ be an NFA for R.
Let $\mathcal{B}$ be an NFA for $L(\mathcal{A}) \overset{\text{max-sdi}}{\longleftarrow} y$, where $y \in L$.

1. Given input w, $\mathcal{B}$ guesses a decomposition $w = x_1 y_1 y_2 y_3 x_2$.
2. $\mathcal{B}$ simulates computation on the prefix $x_1 y_1$, skips y_2, and continues computation on the suffix $y_3 x_2$.
3. $\mathcal{B}$ verifies that the SDI is maximal during the computation.

Decision Problems

- A language L is **SDI-free** with respect to a language R if no word from R can be site-directed inserted into a word from L.
- Similarly, L is **SDI-independent** with respect to R if site-directed insertion of a nonempty word into a word from L does not produce a word from R.

Theorem

Given NFAs $\mathcal{A}$ and $\mathcal{B}$, determining whether $L(\mathcal{A})$ is SDI-free (or SDI-independent) with respect to $L(\mathcal{B})$ is decidable in polynomial time.

Theorem

Given context-free languages L' and R', it is undecidable whether L' is SDI-free (or SDI-independent) with respect to R'.

- Cho et al. proved the following result in an earlier paper.

Proposition

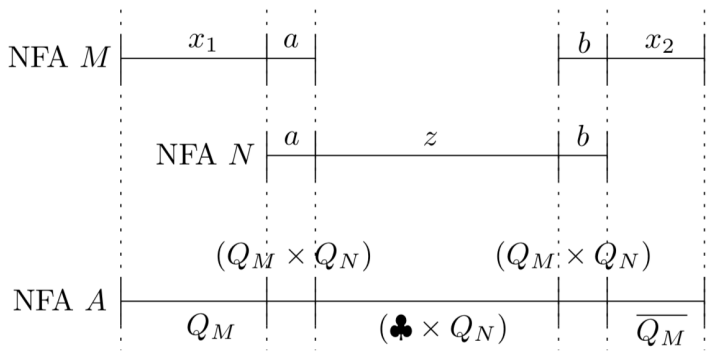
Given NFAs $\mathcal{A}$ and $\mathcal{B}$ with m and n states, respectively, the language $L(\mathcal{A}) \overset{\text{sdi}}{\leftarrow} L(\mathcal{B})$ has an NFA with $3mn + 2m$ states.

- This result gives an upper bound on the nondeterministic state complexity of site-directed insertion.
- It is less obvious what a lower bound might be.
- Likely the bound cannot be improved, but we have no proof.

- **Alphabetic site-directed insertion** is analogous to the SDI operation, except u and v are single symbols.
- For a-SDI, we have more precise nondeterministic state complexity results.

Theorem

Given NFAs $\mathcal{M}$ and $\mathcal{N}$ with m and n states, respectively, the language $L(\mathcal{M}) \xleftarrow{\text{a-sdi}} L(\mathcal{N})$ has an NFA with $mn + 2m$ states.



Theorem

There exist regular languages L_1 and L_2 over a binary alphabet having NFAs with m and n states, respectively, such that any NFA for $L_1 \xleftarrow{\text{a-sdi}} L_2$ requires at least $mn + 2m$ states.

Table of Contents

- We can express SDI as a **semantic shuffle on trajectories** for dealing with language equations.
  - Due to Domaratzki; we use the same definitions/notation here.
- A **trajectory** is a word over $\Gamma = \{0, 1, \sigma\}$.
- The **semantic shuffle** of $x, y \in \Sigma^*$ over a trajectory $t \in \Gamma^*$ is a sequence of instructions that guide the shuffle of x and y.
  - 0: choose the next symbol in x.
  - 1: choose the next symbol in y.
  - σ: synchronized insertion; symbols of x and y must coincide.
- The semantic shuffle of x and y over t is written $x \pitchfork_t y$.

Example

$t = 0\sigma^2 1^4 \sigma^2 0^2$

$x = \texttt{bonjour}$, $y = \texttt{ontrèsjo}$

$x \pitchfork_t y = \texttt{bontrèsjour}$

Proposition

Let $t_{\text{sdi}} = 0^*\sigma^+1^*\sigma^+0^*$. Then, for all languages L_1 and L_2,

$$L_1 \overset{\text{sdi}}{\leftarrow} L_2 = L_1 \pitchfork_{t_{\text{sdi}}} L_2.$$

- We can define $t_{\text{a-sdi}}$ similarly by changing σ^+ to σ.
- Why do we want to model SDI using trajectories?
  - Strong decidability results let us decide linear language equations where the constants are regular languages.
  - We also get regularity-preserving left- and right-inverses.

Theorem

Let L and R be regular languages. Then, for each of the following language equations, it is decidable whether a solution exists:

$$X \overset{\text{sdi}}{\leftarrow} L = R \quad L \overset{\text{sdi}}{\leftarrow} X = R$$

$$X \overset{\text{a-sdi}}{\longleftarrow} L = R \quad L \overset{\text{a-sdi}}{\longleftarrow} X = R$$

- If a solution exists for one of the above equations, we can also find a superset of all solutions for that equation.
- However, the decision algorithm is not polynomial time!
  - The decision procedure uses nondeterministic operations.
  - Requiring complementation blows up the construction.

Table of Contents

Conclusions

- Site-directed insertion, max-SDI, min-SDI, and a-SDI are overlapping variants of the contextual insertion operation.
- SDI is closed under the class of regular languages, but max-SDI and min-SDI are not.
- We can decide in polynomial time whether regular languages are SDI-free or SDI-independent.
- a-SDI has a tight-bounded nondeterministic state complexity of $mn + 2m$.
- We can model SDI as a semantic shuffle to get new decidability results.

Future Work

- Does there exist an algorithm to decide whether a regular language is closed over max-SDI or min-SDI?
  - i.e., Given a regular language L, is $L \overset{\text{max-sdi}}{\longleftarrow} L \subseteq L$ (respectively, $L \overset{\text{min-sdi}}{\longleftarrow} L \subseteq L$)?
- What is the lower bound for the nondeterministic state complexity of SDI?
- Can we solve language equations involving max-SDI or min-SDI?
  - i.e., Given regular languages L and R, can we decide whether $X \overset{\text{max-sdi}}{\longleftarrow} L = R$ (or variants thereof) has a solution?

References

[1] D.-J. Cho, Y.-S. Han, T. Ng, and K. Salomaa. Outfix-guided insertion. *Theor. Comput. Sci.*, 701:70–84, 2017.

[2] D.-J. Cho, Y.-S. Han, K. Salomaa, and T. J. Smith. Site-directed insertion: Decision problems, maximality and minimality. In S. Konstantinidis and G. Pighizzini, editors, *Proc. of DCFS 2018*, volume 10952 of *LNCS*, pages 49–61, Berlin Heidelberg, 2018. Springer-Verlag.

[3] M. Domaratzki. Semantic shuffle on and deletion along trajectories. In C. S. Calude, E. Calude, and M. J. Dinneen, editors, *Proc. of DLT 2004*, volume 3340 of *LNCS*, pages 163–174, Berlin Heidelberg, 2004. Springer-Verlag.

[4] M. Holzer, S. Jakobi, and M. Kutrib. The chop of languages. *Theor. Comput. Sci.*, 682:122–137, 2017.