

Edit Distance

1 The Edit Distance Problem

Definition of a ref-word

Given a finite set of variables $V \subseteq \text{SVars}$ we define the alphabet of ref words as: $\Gamma_V := \{x \vdash, \dashv x\}$. And given an alphabet Σ such that $\Sigma \cap \Gamma_V = \emptyset$ we can define the set of ref words over Σ and V as: $\mathbf{r} \in (\Sigma \cup \Gamma_V)^*$. Next, a ref-word is valid if and only if, every occurrence of a variable in the ref-word is opened exactly once and closed afterwards, exactly once.

Functions on ref-words

We can define the projection of a ref word over a set S , $r \uparrow S$, recursively as:¹

1. $r \in S \rightarrow r \uparrow S = r$
2. $r \notin S \rightarrow r \uparrow S = \epsilon$
3. $(r_1 \cdot r_2) \uparrow S = (r_1 \uparrow S) \cdot (r_2 \uparrow S)$

$\text{Vars}(r)$ is the set of variables $x \in V$ that occurs in the ref-word:

$$\text{Vars}(r) := \{x \in V \mid \exists r_x^{pre}, r_x, r_x^{post} \in (\Sigma \cup \Gamma_V)^* \text{ such that } r = r_x^{pre} \cdot x \vdash \cdot r_x \cdot \dashv x \cdot r_x^{post}\} \quad (1)$$

$\text{tup}(r)$ are the positions each ref-word is referencing, and is defined as:

$$\text{tup}(r) := \{x \mapsto [i_x, j_x] \mid x \in \text{Vars}(R), i_x = |r_x^{pre} \uparrow \Sigma|, j_x = i_x + |r_x \uparrow \Sigma|\} \quad (2)$$

Postulate:

$$\text{valid}(r) \rightarrow |\text{tup}(r)| = |\text{Vars}(r)| \quad (3)$$

Definition of ref-word tuple

Distance between two ref words

Next, given two ref words $r_1, r_2 \in (\Sigma \cup \Gamma_V)^*$ and a distance function $\mathbf{d} : \Sigma^* \times \Sigma^* \rightarrow \mathbb{R}$ the distance $\mathbf{d}_{\mathbf{r}}$ between r_1 and r_2 is defined as: $\mathbf{d}_{\mathbf{r}}(r_1, r_2) = \mathbf{d}(r_1 \uparrow \Sigma, r_2 \uparrow \Sigma)$

Ref-word distance languages

Given a ref-word language $R \subseteq (\Sigma \cup \Gamma_V)^*$ and a distance $k \in \mathbb{R}$, the k-distance ref-word language is defined as:

$$R \pm k = \{r \in (\Sigma \cup \Gamma_V)^* \mid \text{valid}(r), \exists r' (r' \in R \wedge \mathbf{d}_{\mathbf{r}}(r, r') \leq k)\} \quad (4)$$

Given a document d , the spanner over a ref-word language R is:

$$\llbracket R \rrbracket_d = \{\text{tup}(r) \mid r \uparrow \Sigma = d, \exists r \in R\} \quad (5)$$

¹In the paper (Doleschal, 2021) this operation is defined for Σ as $\text{doc}(\sigma)$

Theorem 1. *If the distance function \mathbf{d} is a metric, then:*

$$(R \pm n) \pm k = R \pm (n + k) \quad (6)$$

Proof. Given $n, k \in \mathbb{R}$, we want to prove that: $(R \pm n) \pm k = R \pm (n + k)$. By definition, we have that:

$$(R \pm n) \pm k = \{r \in (\Sigma \cup \Gamma_V)^* \mid \text{valid}(r), \exists r' (r' \in R \pm n \wedge \mathbf{d}_{\mathbf{r}}(r, r') \leq k)\}$$

First we prove that $(R \pm n) \pm k \subseteq R \pm (n + k)$. By contradiction let's assume there exists an element $r_1 \in (R \pm n) \pm k$ such that $r_1 \notin R \pm (n + k)$. By the previous definition, we have that $\exists r' \in R \pm n$ such that $\mathbf{d}_{\mathbf{r}}(r_1, r') \leq k$. By definition of $R \pm (n)$ we have that for any $r \in R$, $\mathbf{d}_{\mathbf{r}}(r', r) \leq n$. Next, by the definition of $R \pm (n + k)$, and our supposition we have that $\mathbf{d}_{\mathbf{r}}(r_1, r) > n + k$ which contradicts the triangle inequality. The proof that $R \pm (n + k) \subseteq (R \pm n) \pm k$ uses this same argument. \square

Variable-set automaton over ref words (VSet-automaton)²

Definition 1. A VSet-automaton is a sextuple $A := (\Sigma, V, Q, q_0, Q_F, \delta)$

- Σ : Alphabet symbols
- V : Finite set of variables
- Q : Finite set of states
- $q_0 \in Q$: Initial state
- $Q_F \subseteq Q$: Set of final states
- $\delta : Q \times (\Sigma \cup \{\epsilon\} \cup \Gamma_V) \rightarrow 2^Q$: Transition function
 - $\Gamma_V := \{x \vdash, \dashv x \mid x \in V\}$
 - 2^Q : power set of Q

Functions over VSet-automaton

$$|A| = |Q| + |Q_F| + |\delta| + 1 \quad (7)$$

$$\text{Vars}(A) := V \quad (8)$$

Ref-word language

The ref-word language of A is: $\mathcal{R}(A) = \mathcal{R}^0(A) = \{r \in \mathcal{L}(A) \subseteq (\Sigma \cup \Gamma_V)^* \mid r \text{ is accepted by the } \epsilon\text{-NFA } A\}$. This is direct from interpreting A as an ϵ -NFA.

Run of a VSet-automaton over a ref-word

Given a ref-word $r = \sigma_1 \cdots \sigma_n$, the run ρ of A is the sequence:

$$\rho := q_0 \xrightarrow{\sigma_1} q_1 \cdots q_{n-1} \xrightarrow{\sigma_n} q_n \quad (9)$$

Where $\forall i \in [0, n) (q_{i+1} \in \delta(q_i, \sigma_{i+1}))$ and $q_n \in Q_F$

From previous publications we know that $r \in \mathcal{R}(A)$ if and only if there is a run ρ of A on r .

²(Doleschal, 2021)

Distance automaton

Given a VSet-automaton A we can define, under Levenshtein distance, the automaton $A \pm 1 := (\Sigma, V, Q', q'_0, Q'_F, \delta')$ Where

- $Q' = \{q_1, \dots, q_{|Q|}\} \cup \{q_1^1, \dots, q_{|Q|}^1\}$ Where there exists two bijective functions:

1. $f : Q \rightarrow \{q_1, \dots, q_{|Q|}\}$
2. $f' : Q \rightarrow \{q_1^1, \dots, q_{|Q|}^1\}$

And two bijective functions F and F' that map f and f' respectively to sets.

- $q'_0 = f(q_0)$
- $Q'_F = Q_F \cup \{q_i^1 \mid q_j \in Q_F \wedge f'(q_j) = q_i^1\}$
- The function δ' is defined by:

$$\delta'(q_i, e) = \begin{cases} F(\delta(f^{-1}(q_i), e)) & e \in \Gamma_V \\ F(\delta(f^{-1}(q_i), e)) \cup q_i^1 \cup \bigcup_{a \in \Sigma} F'(\delta(f^{-1}(q_i), a)) & e.o.c \end{cases}$$

$$\delta'(q_i^1, e) = F(\delta(f^{-1}(q_i), e))$$

Theorem 2. *Given a VSet-automaton with a sequential ref-word language $\mathcal{R}(A)$, using Levenshtein distance we obtain that $\mathcal{R}(A \pm 1) = \mathcal{R}(A) \pm 1$*

Proof. This is equivalent to proving that, given a refword r , $r \in \mathcal{R}(A \pm 1) \leftrightarrow r \in \mathcal{R}(A) \pm 1$. Therefore this is a two part proof.

First let's assume that $r \in \mathcal{R}(A \pm 1)$. In that case we know that there must exist a run ρ of $A \pm 1$ on r . First, by definition of δ' there is no transition from states of the form q_i^1 to q_i . Therefore, there are two cases:

1. There are only states of the form q_i . In this case, the subset of nodes and transitions used have an isomorphism to A using the function f , therefore, $r \in \mathcal{R}(A)$, and therefore $r \in \mathcal{R}(A) \pm 1$
2. There is one transition from a state of the form q_i to q_i^1 . The respective run of $A \pm 1$ on $r = c_1 \dots c_n$ could have one of three forms:

Insertion

$$\rho_{A \pm 1} = f(q_0) \xrightarrow{c_1} f(\phi_1) \dots \xrightarrow{c_{i-1}} f(\phi_i) \xrightarrow{c_i} f'(\phi_i) \xrightarrow{c_{i+1}} f'(\phi_{i+1}) \dots \xrightarrow{c_n} f'(\phi_n) \quad (10)$$

In this case, there exists a run:

$$\rho_{A \pm 1} = q_0 \xrightarrow{c_1} \phi_1 \dots \xrightarrow{c_{i-1}} \phi_i \xrightarrow{c_{i+1}} \phi_{i+1} \dots \xrightarrow{c_n} \phi_n \quad (11)$$

Such that every transition is in A , therefore the word $r' = c_1 \dots c_{i-1} \cdot c_{i+1} \dots c_n$ is in $\mathcal{R}(A)$. And therefore by Levenshtein distance we obtain that $\mathbf{d}_\Gamma(r', r) = 1$. And therefore, $r \in \mathcal{R}(A) \pm 1$

Substitution

TODO

Elimination

TODO

Next, let's assume that $r \in \mathcal{R}(A) \pm 1$. In that case, r is valid, and there exists $r' \in R$ such that using Levenshtein distance: $\mathbf{d}_\Gamma(r, r') \leq 1$. Since Levenshtein distance is discrete, there are two cases:

1. $\mathbf{d}_\Gamma(r, r') = 0$. Then it is clear that $r \in \mathcal{R}(A \pm 1)$ since a subset of $A \pm 1$ forms an isomorphism with A .
2. $\mathbf{d}_\Gamma(r, r') = 1$. In this case, because of the structure of Levenshtein's distance, there are three possible cases:

- (a) **Insertion** $r = c_1 \cdot \dots \cdot c_i \cdot c_{inserted} \cdot c_{i+1} \cdot \dots \cdot c_n$ And $r' = c_1 \cdot \dots \cdot c_n$
 In this case, because there exists a run ρ_A on r' of the form:

$$\rho_A = q_0 \xrightarrow{c_1} \phi_1 \cdots \phi_i \xrightarrow{c_i} \phi_{i+1} \cdots \phi_n^{c_n}$$

Therefore there exists a run $\rho_{A\pm 1}$ on r , and it is:

$$\rho_{A\pm 1} = f(q_0) \xrightarrow{c_1} f(\phi_1) \cdots f(\phi_i) \xrightarrow{c_{ins}} f'(\phi_i) \xrightarrow{c_i} f'(\phi_{i+1}) \cdots \xrightarrow{c_n} f'(\phi_n)$$

- (b) **Elimination** TODO, Same as above
 (c) **Substitution** TODO, Same as above

□