

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}_Σ between r_1 and r_2 is defined as: $\mathbf{d}_\Sigma(r_1, r_2) = \mathbf{d}(r_1 \uparrow \Sigma, r_2 \uparrow \Sigma)$ and the distance \mathbf{d}_Γ is defined extending the distance function \mathbf{b} over $\Sigma \cup \Gamma_V$, treating letters in Γ_V equivalently as words in Σ .

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:

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

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:

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

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 $\mathbf{d}_\Sigma(r_1, r') \leq k$. By definition of $R \pm (n)$ we have that for any $r \in R$, $\mathbf{d}_\Sigma(r', r) \leq n$. Next, by the definition of $R \pm (n + k)$, and our supposition we have that $\mathbf{d}_\Sigma(r_1, r) > n + k$ 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

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 $\epsilon - \text{NFA}$.

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:

$$\rho := q_0 \xrightarrow{\sigma_1} q_1 \cdots q_{n-1} \xrightarrow{\sigma_n} q_n \quad (7)$$

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:

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

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

Definition 2. A ref-word language R is sequential if every ref-word $r \in R$ is valid.

Lemma 1. All runs of $\mathcal{R}(A \pm 1)$ have one of the following structures:

1.

$$\rho_{A \pm 1} = f(q_0) \xrightarrow{c_1} f(\phi_1) \cdots \xrightarrow{c_{a-1}} f(\phi_{a-1}) \xrightarrow{c_a} f'(\phi_{a-1}) \cdots \xrightarrow{c_n} f'(\phi_{n-1})$$

2.

$$\rho_{A \pm 1} = f(q_0) \xrightarrow{c_1} f(\phi_1) \cdots \xrightarrow{c_{a-1}} f(\phi_a) \xrightarrow{c_a} f'(\phi_{a+1}) \xrightarrow{c_{a+1}} f'(\phi_{a+2}) \cdots \xrightarrow{c_n} f'(\phi_n)$$

3.

$$\rho_{A \pm 1} = f(q_0) \xrightarrow{c_1} f(\phi_1) \cdots \xrightarrow{c_{a-1}} f(\phi_a) \xrightarrow{\epsilon} f'(\phi_{a+1}) \xrightarrow{c_a} f'(\phi_{a+2}) \cdots \xrightarrow{c_n} f'(\phi_{n+1})$$

4.

$$\rho_{A \pm 1} = f(q_0) \xrightarrow{c_1} f(\phi_1) \xrightarrow{c_2} \cdots \xrightarrow{c_n} f(\phi_n)$$

Proof. This is direct from the form of the transitions under the definition of $A \pm 1$, and noticing that there is no transition from $f'(q)$ to $f(q)$. \square

Theorem 2. if $\mathcal{R}(A)$ is sequential, then $\mathcal{R}(A \pm 1)$ is sequential.

Proof. For purposes of contradiction let's assume that there exists a ref-word $r = c_1 \cdots c_n \in A \pm 1$ that is not valid. Then there must exist a run $\rho_{A \pm 1}$ on r , and furthermore, by lemma 1, $\rho_{A \pm 1}$ it can have only 4 structures. We will prove by enumeration that all these structures lead to a contradiction.

1. **Insertion:** By definition of the transitions of $A \pm 1$, we have that $c_1 \cdots c_{a-1} \cdot c_{a+1} \cdots c_n \in \mathcal{R}(A)$. This ref-word is valid by the premise of the theorem. Next, c_a cannot be in Γ_V , and therefore we have a contradiction, because variables in all other characters open exactly once, and close exactly once only after they are opened by the definition of a valid ref-word.
2. **Substitution, Elimination** Follow the same argument, in that the word changed is not in Γ_V .
3. **No modification** Trivial.

\square

Theorem 3. *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 ref-word 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 . r is valid due to the theorem 2. By Lemma 1 we have four cases for runs of $\mathcal{R}(a)$, We will now prove by enumeration on these cases:

1. **Insertion:** From the definition of $A \pm 1$, there must be a transition from ϕ_a to ϕ_{a+1} using the letter $a + 1$. Therefore, there is a run:

$$q_0 \xrightarrow{c_1} \dots \xrightarrow{c_{a-1}} \phi_{a-1} \xrightarrow{c_{a+1}} \phi_a \dots \xrightarrow{c_n} \phi_{n-1}$$

And therefore, $c_1 \dots c_a \cdot c_{a+2} \dots c_n \in \mathcal{R}(A)$. And because the extended Levenshtein distance is of 1, and r is valid, then $r \in \mathcal{R}(A) \pm 1$

2. **Substitution, Elimination:** Follow the same argument, changing the sequence into a word and therefore, using the definitions of transitions arriving to a word in $\mathcal{R}(A)$.
3. **No modification:** In this case, we can obtain a run over the same word in $\mathcal{R}(A)$, and therefore, by definition of $\mathcal{R}(A) \pm 1$ we obtain that $r \in \mathcal{R}(A)$

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: $\mathbf{d}_\Sigma(r, r') = \mathbf{d}_\Gamma(r, r') \leq 1$. Since Levenshtein distance is discrete, there are two cases:

1. $\mathbf{d}_\Sigma(r, r') = 0$. 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. $\mathbf{d}_\Sigma(r, r') = 1$. In this case, because of the structure of Levenshtein's distance, there are three possible cases:
 - **Insertion.** In this case the word r looks like: $c_1 \dots c_i \cdot c_{inserted} \cdot c_{i+1} \dots c_n$. Furthermore, $c_{inserted} \in \Sigma$ because if not, $\mathbf{d}_\Gamma(r, r') > \mathbf{d}_\Sigma(r, r')$. Next, since $c_1 \dots c_n \in \mathcal{R}(A)$ this word is accepted by the automaton $A \pm 1$ by the definitions of the transitions.
 - **Substitution / Elimination** Same arguments as Insertion.

□

Theorem 4. *For any sequential automaton A there exists an automaton B such that $\mathcal{R}(B) = \mathcal{R}(A) \pm k$ for all $k \in \mathbb{N}$ under Levenshtein distance.*

Proof. Trivial proof using theorems 1 and ??, and induction over k . □

2 Alternative Definitions

2.1 Navarro 2001

The approximate search problem can therefore be stated as follows. Given a text \mathcal{T} , a regular expression \mathcal{E} , an edit distance between strings $d()$, and a distance threshold k , find all the text positions that start an approximate occurrence of \mathcal{E} in \mathcal{T} , that is, compute the set:

$$\{i, \exists j, \exists S \in L(\mathcal{E}), d(\mathcal{T}i \dots j, S) \leq k\}$$