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 $\Sigma \cap \Gamma_V = \varnothing$ 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 occurance 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, $r \upharpoonright S$, recursively as:[1]

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

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

$$\text{Vars}(r) := \left\{x \in V \;\middle|\; \exists r_x^{pre}, r_x, r_x^{post} \in (\Sigma \cup \Gamma_V)^* \text{ such that } r = r_x^{pre} \cdot x \vdash \cdot r_x \cdot \dashv x \cdot r_x^{post})\right\} \tag{1}$$

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

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

Postulate:

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

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 $\mathbf{d_\Gamma}$ between r_1 and r_2 is defined as: $\mathbf{d_\Gamma}(r_1, r_2) = \mathbf{d}(r_1 \upharpoonright \Sigma, r_2 \upharpoonright \Sigma)$

Ref-word distance languages

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

$$[\![L_{ref}]\!]_{ref}^k = \{r \in (\Sigma \cup \Gamma_V)^* \mid \text{valid(r)}, \exists r'(r' \in L_{ref} \wedge \mathbf{d_\Gamma}(r, r') \leq k)\} \tag{4}$$

Given a document d, the k-distance ref-word language is:

$$[\![L_{ref}]\!]_d^k = \{\text{tup}(r) \mid r \upharpoonright \Sigma = d, \exists r' \in L_{ref}(\mathbf{d_\Gamma}(r, r') \leq k)\} \tag{5}$$

Postulates

$$\left[\!\left[[\![A]\!]_{ref}^{k_1}\right]\!\right]_{ref}^{k_2} = [\![A]\!]_{ref}^{k_1+k_2} \tag{6}$$

$$\left[\!\left[[\![A]\!]_{ref}^{k}\right]\!\right]_{d}^{0} = [\![A]\!]_{d}^{k} \tag{7}$$

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