Distances between languages and reflexivity of relations[1]

Christian Choffrut
LITP, Université Paris 7, 2, Pl. Jussieu
75 251 Paris Cedex 05, France
cc@litp.ibp.fr

Giovanni Pighizzini
Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
via Comelico, 39 – 20135 Milano – Italy
pighizzi@dsi.unimi.it

[1] Please send each communication related to this paper to the second author.

Abstract

We extend the Hamming, edit, prefix, suffix and subword distances between strings to subsets of strings. We show that computing these distances between two rational subsets reduces to computing the weight of an automaton "with distance function" as introduced by Hashiguchi (this latter notion of distance has nothing to do with our notion).

We make a step further by extending the notion of distance between subsets to that of "almost reflexivity" of relations over strings: intuitively a relation is almost reflexive if every element of its domain is in relation with some "close" element in its range and vice versa. Various properties connected to almost reflexivity are investigated. With two exceptions, their decidability status relative to the five notions of distances is settled for the three families of recognizable, synchronous and deterministic relations.

1 Introduction

Comparing strings has been intensively studied during the past two decades, for example in connection with bibliographic search, text editing, alignments of biologic sequences etc. The problem consists usually in associating with two strings a non negative integer which reports how close they are from each other. This usually defines a distance between words, the most popular of which is certainly the one based on the maximal common subsequence, the so-called Levenshtein distance. The basic question concerning the subsequence distance and its extension the edit distance [9] asks whether or not we can design a sequential algorithm that computes the distance in time less than $O(n^2)$ where n is the maximal length of the two words.

Here we tackle the problem of computing the distance between two subsets of words. We assume these subsets to be effectively given. Actually we restrict the effectiveness of the subsets to some types of finite memory device. This is a reasonable assumption since there is no hope of actually computing the distance outside this class as we easily show in this paper.

Let A be a finite alphabet and let A^* be the free monoid it generates. The passage from the distance between strings to the distance between subsets $X, Y \subseteq A^*$ of strings is done in the standard way by resorting to the Hausdorff distance. In particular, whenever X and Y are rational subsets of A^* and whenever we endow the free monoid with the Hamming distance, the problem can be reduced to Hashiguchi's result on automata with distance functions (caveat: the term distance here refers to a different concept).

It is worthwhile observing that some fundamental problems can be stated in terms of distances. For example, asking whether or not the distance between two subsets, whatever the distance, is equal to 0 amounts to asking whether or not they are equal. This question is undecidable for context-free languages, so this gives a limit to our expectations and shows that we are quickly facing difficult problems.

Asserting that two subsets X and Y have finite distance is a claim that the binary relation $X \times Y$ is "almost" reflexive in some precise sense, with respect to the distance under consideration. Thus we were led to consider the more general question of asking whether or not a given relation over A^* is almost reflexive: a relation $R \subseteq A^* \times A^*$ is k–reflexive, for some integer $k \leq \infty$, if every element x of the domain is at distance at most k from some element of the range y, with $(x, y) \in R$, and vice versa. It is almost reflexive if $k < \infty$. J. Johnson [8] proved that reflexivity is undecidable for rational relations. In this work, we strengthen this result by showing that the problem remains undecidable even when we restrict our attention to *deterministic* rational relations. This result is a new contribution to the theory of rational relations and has some important consequences concerning the problems we are interested in in this paper. First of all, we cannot decide whether or not a deterministic rational relation is k–reflexive, for any given integer k, with respect to whatever distance. Moreover, in the case of the distances considered in this work, namely, the Hamming, prefix, suffix, subword and edit distances, this problem remains undecidable even we are given an oracle asserting that the relation is almost reflexive. In addition, we prove that it is undecidable whether or not a deterministic rational relation is almost reflexive.

Subsequently, in order to deal with the largest classes of relations for which the previously mentioned properties are decidable, we restrict our attention to proper subfamilies of deterministic rational relations, in particular that of synchronized rational relations [4] and the smaller class of recognizable relations.

We show that Hashiguchi's above mentioned problem is very close to the problem of deciding whether or not a synchronized rational relation is almost reflexive with respect to the Hamming distance. In fact, we show that each one of these two problems can be linearly reduced to one another. As a consequence, we prove that the smallest $k \leq \infty$, such that a synchronized rational relation is k–reflexive with respect to the Hamming distance can be effectively computed.

In the case of the subword and of the edit distance, the largest class for which we show that this computation can be done is that of recognizable relations. This result does not seem to be extendable to synchronized rational relations. However, we are able to prove that given an oracle asserting that a synchronized rational relation is almost reflexive, then it is decidable, given an integer k, whether or not it is k–reflexive with respect to the subword or edit distance. These results are summarized in Table 1 and Table 2.

We conclude the paper by studying the unary case, namely the case of relations defined over a one letter alphabet. In this case, the family of rational relations is a Boolean algebra. Making use of the correspondence between rational relations over a one letter alphabet and formulae in the Presburger arithmetic [5], we are able to prove that all the considered problems are solvable.

	Rec	Syn	$DRat$
Is R k–reflexive?	S	S	U
Is R almost reflexive?	S	S	U
Find the smallest k s.t. R is k–reflexive	S	S	U
Given R almost reflexive, find the smallest k s.t. R is k–reflexive	S	S	U

S=Solvable, U=Unsolvable
Rec=Recognizable relations
Syn=Synchronized rational relations
$DRat$=Deterministic rational relations

Table 1: k–reflexivity with respect to the Hamming distance.

2 Preliminaries

Given a finite alphabet A, we denote by A^* the free monoid of strings over A, and by ϵ the empty string. Given a string $x \in A^*$, we denote by $|x|$ its length and by x_i the ith symbol of x, $i = 1, \ldots, |x|$. We recall that a string $z \in A^*$ is a *subword* or *factor* of x if and only if there exist two strings $u, v \in A^*$ such that $x = uzv$. If $u = \epsilon$ ($v = \epsilon$, respectively) then

	Rec	*Syn*	*DRat*
Is R k–reflexive?	S	S	U
Is R almost reflexive?	S	?	U
Find the smallest k s.t. R is k–reflexive	S	?	U
Given R almost reflexive, find the smallest k s.t. R is k–reflexive	S	S	U

Table 2: k–reflexivity with respect to the prefix, suffix, subword and edit distances.

z is a *prefix* (*suffix*, resp.) of x.

2.1 Distances between strings – Hausdorff distances

A distance over A^* is a function $d : A^* \times A^* \to \mathbb{N} \cup \{\infty\}$ such that, for all $x, y, z \in A^*$ the following holds

i) $d(x, y) = 0$ if and only if $x = y$,

ii) $d(x, y) = d(y, x)$, and

iii) $d(x, y) \leq d(x, z) + d(z, y)$.

Strings over A^* can be compared symbol by symbol. In particular, one can define the *Hamming distance* between x and y as the number of positions where the symbol of x is different from the symbol of y. More formally:

Definition 2.1 *Given $x, y \in A^*$, the* Hamming distance *between x and y, denoted by $d_H(x, y)$ is defined as:*

$$d_H(x, y) = \begin{cases} \#\{i \mid x_i \neq y_i\} & \text{if } |x| = |y| \\ \infty & \text{otherwise.} \end{cases}$$

Furthermore, two strings can be compared via their longest common prefix, suffix and subword, respectively. This leads to the following notions:

Definition 2.2 *Given $x, y \in A^*$, the* prefix, suffix *and* subword distance *between x and y, denoted respectively by $d_p(x, y)$, $d_s(x, y)$ and $d_f(x, y)$, are defined as:*

$$\begin{aligned} d_p(x, y) &= |x| + |y| - 2 \max \{|z| \mid x, y \in zA^*\} \\ d_s(x, y) &= |x| + |y| - 2 \max \{|z| \mid x, y \in A^*z\} \\ d_f(x, y) &= |x| + |y| - 2 \max \{|z| \mid x, y \in A^*zA^*\} \end{aligned}$$

Note that, for $x, y \in A^*$,

$$d_s(x,y) = d_p(x^R, y^R) \tag{1}$$

where, given a string $x = a_1 a_2 \ldots a_n$, x^R denotes its mirror image, namely the string $a_n \ldots a_2 a_1$.

Another notion of distance between strings is defined by considering the *edit operations* of insertion, deletion, and substitution, which transform one string into another. More precisely, we can consider the binary relation $\vdash$ over A^* by setting for all $x, y \in A^*$, $x \vdash y$ if and only if there exist $u, v \in A^*$, $a, b \in A$, with $a \neq b$, such that either condition is satisfied

i) $x = uav$, $y = ubv$ (substitution)

ii) $x = uav$, $y = uv$ (deletion)

iii) $x = uv$, $y = ubv$ (insertion)

Note that $\vdash$ is symmetric. As usual, we denote by $\overset{k}{\vdash}$ the composition of $\vdash$ with itself, k times.

Definition 2.3 *Given two strings $x, y \in A^*$, the* edit distance $d_e(x,y)$ *between x and y is the minimum number of edit operations which transform x into y, i.e.,*

$$d_e(x,y) = \min \{k \mid x \overset{k}{\vdash} y\}.$$

There exists a more amenable definition of the edit distance that uses the notion of *alignments*, cf. [9]. Indeed, saying that x and y have edit distance k means that they can be written as $x = x_0 a_1 x_1 a_2 \ldots a_n x_n$, and $y = y_0 b_1 y_1 b_2 \ldots b_n y_n$ where the a_i's and the b_i's are letters, the x_i's and y_i's are words and

$$k = |x_0 x_1 \ldots x_n| + |y_0 y_1 \ldots y_n| + \#\{i \mid a_i \neq b_i\}.$$

As a consequence, for all strings $x, y \in A^*$, and all factorizations $x = x_1 \ldots x_n$, $(n > 0)$ the following equality holds:

$$d_e(x,y) = \min \{d_e(x_1, y_1) + \ldots + d_e(x_n, y_n) \mid y = y_1 \ldots y_n\}. \tag{2}$$

Each distance over A^* can be extended to a distance between strings and languages from one side and between languages from the other. The distance between a string x and a language L is defined as the minimum of the distances between x and the strings belonging to L. Thus

Definition 2.4 *Let A be a finite alphabet, $d : A^* \times A^* \to \mathbb{N} \cup \{\infty\}$ a distance, $x \in A^*$ a string, and $L \subseteq A^*$ a language.*

The distance *between x and L is defined as:*

$$d(x,L) = \begin{cases} \min \{d(x,y) \mid y \in L\} & \textit{if } L \neq \emptyset \\ \infty & \textit{otherwise.} \end{cases}$$

The distance between two languages can be defined in a standard way, by resorting to the Hausdorff distance. More precisely:

Definition 2.5 *Let A be a finite alphabet, $d : A^* \times A^* \to \mathbb{N} \cup \{\infty\}$ a distance and $L', L'' \subseteq A^*$.*

The relative distance *from L' to L'' is defined as:*

$$d(L'|L'') = \sup\{d(x, L'') \mid x \in L'\}.$$

The distance *between L' and L'' is defined as:*

$$d(L', L'') = \max\{d(L'|L''), d(L''|L')\}.$$

Recalling that the equality of context–free languages is undecidable, it is immediate to conclude that the distance between context–free languages is not computable. So, it is natural to restrict our attention to the family of rational languages.

2.2 Relations over strings – Almost reflexivity

The notion of distance between two languages can be generalized to that of almost reflexivity of relations. To this aim, we now recall some basic definitions concerning binary relations over strings, and then we introduce the notion of k–reflexivity.

Given a binary relation $R \subseteq A^* \times B^*$, we denote by R^{-1} the relation $\{(y, x) \in B^* \times A^* \mid (x, y) \in R\}$. The *domain* and the *range* of R are denoted by $Dom(R)$ and $Ran(R)$ respectively, namely, $Dom(R) = \{x \mid \exists y \text{ s.t. } (x, y) \in R\}$, and $Ran(R) = Dom(R^{-1})$.

Let us slightly modify the usual definition of reflexivity. For all $L \subseteq A^*$ set $i_L = \{(x, x) \mid x \in L\}$. Then a relation $R \subseteq A^* \times A^*$ is *reflexive* if and only if, for all $x \in Dom(R) \cup Ran(R)$, the pair (x, x) belongs to R, namely, $i_{Dom(R) \cup Ran(R)} \subseteq R$. Observe that this implies $Dom(R) = Ran(R)$.

We now introduce the *k–reflexivity* with respect to a given distance d over A^*. Intuitively, a relation $R \subseteq A^* \times A^*$ is k–reflexive, for some $k \in \mathbb{N} \cup \{\infty\}$, if and only if every element of its domain is related by R to some element of its range at distance at most k, and vice versa. Formally:

Definition 2.6 *Given a distance d over A^*, a relation $R \subseteq A^* \times A^*$, and an element $k \in \mathbb{N} \cup \{\infty\}$, the relation R is k–*reflexive *with respect to d, if and only if both the following statements hold:*

i) $\forall x \in Dom(R)\ \exists y \in A^*$ *s.t.* $(x, y) \in R$ *and* $d(x, y) \leq k$,

ii) $\forall y \in Ran(R)\ \exists x \in A^*$ *s.t.* $(x, y) \in R$ *and* $d(x, y) \leq k$.

The relation R is almost reflexive *(with respect to d) if and only if it is k–reflexive for some $k < \infty$.*

Note that, for any distance d, a relation is reflexive if and only if it is 0–reflexive. Furthermore, the notion of k–reflexivity of relations is, indeed, a generalization of that of distance between languages. In fact, the distance between two languages L' and L'' is bounded by k if and only if the relation $L' \times L''$ is k–reflexive. Hence, L' and L'' have finite distance if and only if the relation $L' \times L''$ is almost reflexive.

We now recall the definition of the main classes of relations we are interested in. As usual, given a monoid M, we denote by $Rec(M)$ and $Rat(M)$ the set of recognizable and rational subsets of M, respectively.

The class $Rat(A^* \times B^*)$ of *rational relations* over $A^* \times B^*$ can be characterized using rational transducers.

Definition 2.7 *A* rational transducer $\mathcal{T}$ *is a tuple* (A, B, Q, I, F, E) *where:*

i) A, B *are two finite alphabets,*

ii) Q *is the finite set of states,*

iii) $I \subseteq Q$ *is the set of initial states,*

iv) $F \subseteq Q$ *is the set of final states,*

v) $E \subseteq Q \times A^* \times B^* \times Q$ *is the finite set of transitions.*

A *path* from q_0 to q_n in $\mathcal{T}$, $q_0, q_n \in Q$, is a sequence of transitions of E of the form

$$(q_0, u_1, v_1, q_1)(q_1, u_2, v_2, q_2) \dots (q_{n-1}, u_n, v_n, q_n) \tag{3}$$

The path is said to be a *successful path* if and only if $q_0 \in I$ and $q_n \in F$. The *label* of the path is the componentwise concatenation of labels, namely, the pair (u, v) where $u = u_1 u_2 \dots u_n$ and $v = v_1 v_2 \dots v_n$. It can also be convenient to write the path as

$$q_0 \xrightarrow{(u_1,v_1)} q_1 \xrightarrow{(u_2,v_2)} q_2 \dots q_{n-1} \xrightarrow{(u_n,v_n)} q_n$$

or for short as

$$q_0 \xrightarrow{(u,v)} q_n$$

The *relation* $R \in A^* \times B^*$ accepted by the transducer $\mathcal{T}$ is the set of labels of successful paths of $\mathcal{T}$.

Without loss of generality, we can suppose that the set E of transitions of the transducer $\mathcal{T}$ is a subset of $(Q \times A \times \{\epsilon\} \times Q) \cup (Q \times \{\epsilon\} \times B \times Q)$.

There exists a deterministic version of the previous notion (also knwon as *deterministic 2–automaton*, cf. [10]).

Definition 2.8 *A* deterministic transducer $\mathcal{A}$ *is a tuple* (A, B, Q, I, F, E) *where* A*,* B*,* Q*,* I*,* F *are defined as for general transducers, and:*

i) I *is reduced to one element;*

ii) $E \subseteq (Q \times (A \cup \{\not{c}\}) \times \{\epsilon\} \times Q) \cup (Q \times \{\epsilon\} \times (B \cup \{\not{c}\}) \times Q)$*, where* $\not{c} \notin A \cup B$*;*

iii) $(q, a, b, p'), (q, a, b, p'') \in E$ *implies* $p' = p''$*;*

iv) $(q, a, b, p'), (q, c, d, p'') \in E$ *implies* $a = c = \epsilon$ *or* $b = d = \epsilon$*.*

The relation accepted by $\mathcal{A}$ is the set of all pairs $(u, v) \in A^* \times B^*$ such that $(u\not{c}, v\not{c})$ is the label of a successful path of $\mathcal{A}$. The class $DRat(A^* \times B^*)$ of *deterministic rational relations* over $A^* \times B^*$ is the set of all relations accepted by deterministic transducers and it is a proper subclass of $Rat(A^* \times B^*)$ [10]. It is not difficult to see that $DRat(A^* \times B^*)$ is closed under complementation.

The notion of *synchronized transducer* is a specialization of the previous ones. Intuitively, a transducer $\mathcal{T}$ is synchronized if and only if each path in $\mathcal{T}$ can be divided in two (possibly empty) parts. The first part contains only "letter–to–letter" transitions, i.e., transitions (q, a, b, p) such that $a \in A$, $b \in B$. In the second part, when all symbols from the shortest component are exhausted, only transitions from the other component are performed.

Definition 2.9 *A transducer $\mathcal{T}$ is* synchronized *if and only if the sets Q and E can be partitioned into three subsets Q', Q_A, and Q_B, E', E_A, and E_B, respectively, such that:*

$$\begin{array}{l} E' \subseteq Q' \times A \times B \times Q \\ E_A \subseteq Q_A \times A \times \{\epsilon\} \times Q_A \\ E_B \subseteq Q_B \times \{\epsilon\} \times B \times Q_B \end{array} \tag{4}$$

The class $Syn(A^* \times B^*)$ of *synchronized rational relations* over $A^* \times B^*$ [3] (see also [4]) is defined as the class of relations accepted by synchronized rational transducers.

The following proper inclusions, where A and B are assumed non empty, are well known (where $Rat(A^*) \times Rat(B^*)$ stands for the family of subsets of the form $X \times Y$ with $X \in Rat(A^*)$ and $Y \in Rat(B^*)$) [10, 2, 8, 4]:

$$Rat(A^*) \times Rat(B^*) \subset Rec(A^* \times B^*) \subset Syn(A^* \times B^*) \subset DRat(A^* \times B^*) \subset Rat(A^* \times B^*)$$

We conclude this section with automata known as automata with a distance function cf. [6, 7], but we prefer to call them weighted automata in order to avoid any confusion.

Definition 2.10 *A* weighted automaton *$\mathcal{A}$ is a quintuple (A, Q, ω, I, F), where A is an input alphabet, Q is the finite set of states, I is the set of initial states, F is the set of final states and $\omega : Q \times A \times Q \to \mathbb{N} \cup \{\infty\}$ is a* weight function.

Similarly to ordinary automata, a path is a sequence of quadruples of the form

$$q_0 \xrightarrow{a_1/d_1} q_1 \xrightarrow{a_2/d_2} q_2 \ldots q_{n-1} \xrightarrow{a_n/d_n} q_n$$

where $d_i = \omega(q_{i-1}, a_i, q_i)$ for $i = 1, \ldots, n$. The weight of the path is $\sum_{i=1,\ldots,n} d_i$. Intuitively, the weight of an input string is the minimum of the weights of the paths labeled by that string. Formally, we extend the weight function ω to strings, by setting, for $q, p \in Q$, $x \in A^*$, $a \in A$:

i) $\omega(q, \epsilon, p) = \begin{cases} 0 & \text{if } q = p \\ \infty & \text{otherwise;} \end{cases}$

ii) $\omega(q, xa, p) = \min \{\omega(q, x, r) + \omega(r, a, p) \mid r \in Q\}$;

iii) $\omega(x) = \min \{\omega(q, x, p) \mid q \in I, p \in F\}$.

Finally, we can define the *weight* of $\mathcal{A}$, denoted by $\omega(\mathcal{A})$, as the maximum of the weights of all the strings accepted by $\mathcal{A}$, namely:

$$\omega(\mathcal{A}) = \sup \{\omega(x) \mid x \in L(\mathcal{A})\}.$$

By convention, if $\mathcal{A}$ accepts the empty set, then $\omega(\mathcal{A}) = \infty$.

The *underlying automaton* is the finite (non necessarily deterministic) automaton (A, Q, δ, I, F) where the transition function $\delta : Q \times A \to 2^Q$ satisfies for all $q, p \in Q$ and $a \in A$, $p \in \delta(q, a)$ if and only if $\omega(q, a, p) < \infty$.

The following result was proved by Hashiguchi in [6, 7].

Theorem 2.1 *Given an automaton $\mathcal{A}$ with a weight function ω, the weight $\omega(\mathcal{A})$ is computable.*

3 Undecidability results

In this section, we study the problem of deciding whether or not a relation is k–reflexive, for a fixed k, without specifying the distance. More precisely, we show that for deterministic rational relations, this problem and other related problems are undecidable.

The results of this section are mainly based on the following theorem, which extends to deterministic rational relations a result proven in [8] for rational relations:

Theorem 3.1 *It is undecidable whether or not a deterministic rational relation $R \subseteq A^* \times A^*$ is reflexive, i.e., whether or not $i_{Dom(R) \cup Ran(R)} \subseteq R$ holds.*

Proof. Given a deterministic rational relation $R \subseteq A^* \times A^*$, let $X = A^* - Dom(R)$ and $Y = A^* - Ran(R)$. Since $X \times Y$ is a recognizable relation, the relation $R \cup (X \times Y)$ is deterministic too. Moreover, R is reflexive if and only if the diagonal i_{A^*} is contained in $R \cup (X \times Y)$.

Let $u_1, \ldots, u_p, v_1, \ldots, v_p \in A^*$ be an instance of Post's Correspondence Problem (PCP) such that the sets $\{u_1, \ldots, u_p\}$ and $\{v_1, \ldots, v_p\}$ are prefix codes. As shown in [11], PCP remains undecidable even under this hypothesis. Now, consider the following relation:

$$R = \{(u_i, v_i) \mid i = 1, \ldots, n\}^+$$

It is not difficult to see that $R \in DRat(A^* \times A^*)$. Moreover, since $DRat(A^* \times A^*)$ is closed under complementation, R^c belongs to $DRat(A^* \times A^*)$. We observe that R^c contains the diagonal i_{A^*} if and only if the given instance of PCP has no solution which completes the proof. ∎

As an immediate consequence, for an arbitrary distance d, it is undecidable, given a deterministic rational relation R and an integer $k \geq 0$, whether or not R is k–reflexive. Actually, in the case of the Hamming, subword and edit distances, this observation can be strengthened in two directions. On one hand, it is undecidable whether or not a deterministic rational relation is almost reflexive (we do not require the bound k, we just inquire about the existence of a finite bound). On the other hand, the problem of deciding, given a deterministic rational relation and an integer k, whether or not it is k–reflexive, remains undecidable even knowing that R is almost reflexive. This is what we turn to prove, but before we need to establish a technical lemma.

Lemma 3.1 *Consider the Hamming (resp. prefix, suffix, subword, edit) distance over A^*. Given $R \subseteq A^* \times A^*$, let $\$ \notin A$, and define $R_\$ = \{(u\$, v\$) \mid (u, v) \in R\}$. Then the following conditions are equivalent*

i) *R is reflexive,*

ii) *$R_\$^*$ is almost reflexive,*

iii) $R_\$^*$ *is reflexive.*

Proof. The implications i) $\Rightarrow$ iii) and iii) $\Rightarrow$ ii) are trival.

ii) $\Rightarrow$ i) Suppose that $R_\$^*$ is k–reflexive for some integer $k < \infty$ and fix $u \in Dom(R)$. For $x = (u\$)^{k+1}$, there exists $y \in (A \cup \{\$\})^*$ such that $(x, y) \in R_\$^*$ and $d(x, y) \leq k$.

Let us first deal with the Hamming and the edit distances. The following holds (trivially for the Hamming distance and as a consequence of equality (2) for the edit distance). The string y can be decomposed as the product of $k+1$ factors $y_0, y_1, \ldots, y_k$ such that

$$d(x, y) = d(u\$, y_0) + d(u\$, y_1) + \ldots + d(u\$, y_k)$$

This implies that $y_i = u\$$, for some i, $0 \leq i \leq k$. Then $(u\$, u\$) \in R_\$$ and $(u, u) \in R$.

In the case of the prefix (suffix, subword, resp.) distance, the distance is $|x| + |y| - 2l(x, y) \leq k$, where $l(x, y)$ denotes the length of the longest common prefix (suffix, subword, resp.) of x and y. So,

$$2l(x, y) \geq |x| + |y| - k \geq (|u| + 1)(k + 1) + l(x, y) - k = (k + 1)|u| + 1 + l(x, y)$$

and then

$$l(x, y) \geq (k + 1)|u| + 1.$$

For the prefix distance, this implies that x and y have a common prefix of length $|u|+1$, namely $u\$$ is a common prefix of x and y. Then, by construction, we can conclude that $(u\$, u\$) \in R_\$$, and $(u, u) \in R$.

The case of the suffix distance is similar.

In the case of the subword distance, from the last inequality it turns out that, for $k \geq 1$, x and y have a common subword of length $2|u| + 1$. Then, by construction, we can conclude that $(u\$, u\$) \in R_\$$, and $(u, u) \in R$. For $k = 0$ we immediately obtain $x = y = u\$$. Thus $(u\$, u\$) \in R_\$^*$, and $(u, u) \in R$.

In a similar way, given $v \in Ran(R)$, it is possible to prove that $(v, v) \in R$. Hence, we have shown that if $R_\$^*$ is k–reflexive then R is reflexive. ∎

Theorem 3.2 *It is undecidable, given a deterministic rational relation, whether or not is almost reflexive, with respect to the Hamming, prefix, suffix, subword, and edit distance, respectively.*

Proof. Given $R \in DRat(A^* \times A^*)$, let $\$ \notin A$, and consider the relation

$$R_\$ = \{(u\$, v\$) \mid (u, v) \in R\},$$

and its Kleene closure $R_\$^*$. Of course, $R_\$$ and $R_\$^*$ are deterministic rational relations. By Theorem 3.1, reflexivity of deterministic rational relations is undecidable. We can conclude via Lemma 3.1. ∎

The following is a more precise statement than just saying that k–reflexivity is undecidable, since the integer k is not part of the input. For $k = 0$ this is Theorem 3.1.

Lemma 3.2 *Let k be an integer. Given a deterministic rational relation $R \subseteq A^* \times A^*$ it is undecidable whether or not it is k–reflexive relative to the Hamming, prefix, suffix, subword and edit distance respectively.*

Proof. From R construct the relation $R_\$$ as in Lemma 3.1. Let $\alpha, \beta \notin A \cup \{\$\}$. Define two words $x, y \in \{\alpha, \beta\}^*$ of length greater than or equal to k, such that $d(x, y) = k$. The subset $S = {R_\$}^* \cup \{(x, y)\}$ is clearly a deterministic rational relation. By Lemma 3.1, R is reflexive if and only if S is k–reflexive, which completes the proof. ∎

Finally, we show that knowing that a relation is k–reflexive does not help deciding whether or not it is $(k-1)$–reflexive.

Theorem 3.3 *Let $k > 0$ be an integer. It is undecidable, given a k–reflexive deterministic rational relation $R \in DRat(A^* \times A^*)$ whether or not it is $(k-1)$–reflexive, with respect to Hamming, prefix, suffix, subword, and edit distance, respectively.*

Observe that this result requires some hypothesis on the distance. Indeed, it does not hold for $k > 1$ and the distance defined by $d(x, y) = 1$ iff $x \neq y$.

Proof. There is no loss of generality in the cases of the Hamming and the edit distances to assume that all words in $Dom(R) \cup Ran(R)$ have length at least k. Indeed, let $a \in A$ be an arbitrary letter and set $R' = \{(a^k, a^k)\}R$. Then for all k, R' is k–reflexive if and only if R is k–reflexive.

Let $\sharp \notin A$ and for every $u = u_1 u_2 \ldots u_n$, $u_i \in A$, $i = 1, \ldots, n$, let $u^\sharp$ denote the word $\sharp^k u_{k+1} \ldots u_n$. Now consider

$$R^\sharp = R \cup R_0 \cup R_1 \cup R_2$$

where for the Hamming and edit distances we set

$$\begin{aligned} R_0 &= \{(u, u^\sharp) \mid u \in Dom(R)\} \\ R_1 &= \{(u^\sharp, u) \mid u \in Ran(R)\} \\ R_2 &= \{(u^\sharp, u^\sharp) \mid u \in Dom(R) \cup Ran(R)\} \end{aligned}$$

for the prefix and subword distances we set

$$\begin{aligned} R_0 &= \{(u, u\sharp^k) \mid u \in Dom(R)\} \\ R_1 &= \{(u\sharp^k, u) \mid u \in Ran(R)\} \\ R_2 &= \{(u\sharp^k, u\sharp^k) \mid u \in Dom(R) \cup Ran(R)\} \end{aligned}$$

and for the suffix distance we reduce to prefix distance via (1).

It is not difficult to observe that in the three cases, $R^\sharp$ is a k–reflexive deterministic rational relation. Furthermore, $R^\sharp$ is $(k-1)$–reflexive if and only if R is $(k-1)$–reflexive. The result follows from Lemma 3.2. ∎

4 k–reflexivity with respect to the Hamming distance

As a consequence of Section 3, we know that the problem of deciding whether or not a deterministic rational relation is k–reflexive with respect to the Hamming distance is unsolvable. Now, we prove that this problem turns out to be solvable when we restrict our attention to the subclass of synchronized rational relations.

Theorem 4.1 *For any synchronized rational relation R, the smallest $k \leq \infty$, such that R is k–reflexive with respect to the Hamming distance is effectively computable.*

Proof. Set $K = (A \times A)^*$. Then $R \cap K$ is a synchronized rational relation because the class of synchronized rational relations is closed under the Boolean operations. If $Dom(R \cap K) \subset Dom(R)$ or $Ran(R \cap K) \subset Ran(R)$ then the Hamming distance is infinite, else for all k, R is k–reflexive if and only if $R \cap K$ is k–reflexive. Since all these properties are effective, we assume without loss of generality that $R \subseteq K$, i.e., it can be realized by a synchronized transducer $\mathcal{T} = (A, A, Q, I, F, E)$ for which $E \subseteq Q \times A \times A \times Q$ holds. Clearly, R is a synchronized rational relation if and only if so is R^{-1}. Because of Definition 2.6, it thus suffices to prove that there exists a weighted automaton $\mathcal{A} = (A, Q, \omega, I, F)$ satisfying $\omega(x) = \min \{d_H(x, y) \mid (x, y) \in R\}$ for all $x \in Dom(R)$.

Intuitively, the transitions of the automaton $\mathcal{A}$ are obtained by "forgetting" the third component of the $\mathcal{T}$-transitions, and the weight function ω is defined by considering matches between the second and third components of the $\mathcal{T}$-transitions. Formally, the weight function ω satisfies

$$\omega(q, a, p) = \begin{cases} 0 & \text{if } (q, a, a, p) \in E \\ 1 & \text{if } (q, a, a, p) \notin E \text{ and } \exists b \in A \text{ s.t. } (q, a, b, p) \in E \\ \infty & \text{otherwise.} \end{cases}$$

Consider a fixed $x \in A^*$.

Let $y \in A^*$ satisfy $(x, y) \in R$ and put $x = a_1 a_2 \ldots a_n$ and $y = b_1 b_2 \ldots b_n$ where the a_i's and the b_i's are letters. There exists a successful path in $\mathcal{T}$

$$q_0 \xrightarrow{(a_1,b_1)} q_1 \xrightarrow{(a_2,b_2)} q_2 \ldots q_{n-1} \xrightarrow{(a_n,b_n)} q_n$$

For $i = 1, \ldots, n$, set $d_i = \omega(q_{i-1}, a_i, q_i)$. By definition of ω we have $d_i \leq d_H(a_i, b_i)$. The previous path yields a successful path in $\mathcal{A}$

$$q_0 \xrightarrow{a_1/d_1} q_1 \xrightarrow{a_2/d_2} q_2 \ldots q_{n-1} \xrightarrow{a_n/d_n} q_n$$

which implies $\omega(x) \leq \#\{i \mid a_i \neq b_i\} = d_H(x, y)$. Conversely, consider a successful path in $\mathcal{A}$ with weight $\omega(x)$

$$q_0 \xrightarrow{a_1/d_1} q_1 \xrightarrow{a_2/d_2} q_2 \ldots q_{n-1} \xrightarrow{a_n/d_n} q_n$$

By definition of ω, for $i = 1, \ldots, n$ we have $(q_{i-1}, a_i, c_i, q_i) \in E$ where $c_i = a_i$ if and only if $d_i = 0$. This yields a successful path in $\mathcal{T}$

$$q_0 \xrightarrow{(a_1,c_1)} q_1 \xrightarrow{(a_2,c_2)} q_2 \ldots q_{n-1} \xrightarrow{(a_n,c_n)} q_n$$

showing that $d_H(x, z) \leq \omega(x)$ where $z = c_1 c_2 \ldots c_n$. This completes the proof. ∎

The problems of deciding whether or not a synchronized rational relation is k–reflexive with respect to the Hamming distance and the problem of computing weights of automata are closely related. Indeed, as shown in the proof of Theorem 4.1, the problem of deciding whether or not a synchronized rational relation is k–reflexive with respect to the Hamming distance reduces to the problem of computing the weights of two automata with suitable weight functions. The converse also holds.

Theorem 4.2 *The problems of deciding whether or not the weight of a weighted automaton is bounded and that of deciding whether or not a synchronized rational relation is almost reflexive are linearly Turing reducible to each other.*

Proof. The previous theorem gives a linear reduction of the second to the first problem. Let us show the converse reduction.

Without loss of generality we may assume $\omega(q, a, p) \in \{0, 1, \infty\}$ for any $q, p \in Q$, $a \in A$ (define $\omega'(q, a, p) = 1$ if $1 < \omega(q, a, p) < \infty$ and $\omega'(q, a, p) = \omega(q, a, p)$ otherwise). Now we construct a relation R which is k–reflexive if and only if $\omega(\mathcal{A})$ is bounded by k.

Let $\sharp$ be a symbol not belonging to A and let $\mathcal{A} = (A, Q, \omega, I, F)$ be the weighted automaton. With the notations of equation (4), we define the synchronized transducer satisfying

i) $Q' = Q$

ii) $(q, a, b, p) \in E'$ if and only if $\begin{cases} a = b & \text{and} \quad \omega(q, a, p) = 0 \\ \text{or} & \\ a \in A, b = \sharp & \text{and} \quad \omega(q, a, p) = 1 \end{cases}$

Denoting by R the synchronized rational relation realized by the previous automaton, the relation $R \cup R^{-1} \cup i_{Ran(R)}$ is synchronized too and satisfies

$$\omega(\mathcal{A}) \leq k \text{ if and only if } R \cup R^{-1} \cup i_{Ran(R)} \text{ is } k\text{–reflexive.}$$

This completes the proof. ∎

5 k–reflexivity with respect to the edit distance

In this section we study the problem of deciding whether or not a relation is k–reflexive with respect to the edit distance. In particular, we show that this problem is solvable for recognizable relations. As in the case of the Hamming distance and of synchronized relations, this result is obtained by a reduction to the problem of Hashiguchi.

We start with a general result that intuitively asserts that k-reflexivity for recognizable relations reduces to relative distances between rational subsets.

Lemma 5.1 *Let d be a distance over A^* for which given $X, Y \in Rat(A^*)$, $d(X|Y)$ can be effectively computed. Then given a recognizable subset $R \subseteq A^* \times A^*$, the minimum integer k for which R is k–reflexive is computable.*

Proof. By Mezei's Theorem [1, Thm. III. 1. 5] any recognizable relations $R \subseteq A^* \times A^*$ is a finite union of the form $R = \bigcup_{i=1}^{m} L'_i \times L''_i$ where $m \geq 0$ and $L'_1, \ldots, L'_m, L''_1, \ldots, L''_m$ are regular languages. Note that, for $i, j = 1, \ldots, m$:

$$(L'_i \times L''_i) \cup (L'_j \times L''_j) = ((L'_i - L'_j) \times L''_i) \cup ((L'_j - L'_i) \times L''_j) \cup ((L'_i \cap L'_j) \times (L''_i \cup L''_j))$$

So, we can express R as $R = \bigcup_{i=1}^{p} M'_i \times M''_i$ where the M'_i's and the M''_i's are regular languages of A^*, the M'_i's are pairwise disjoint and $M'_i \neq \emptyset$, $M''_i \neq \emptyset$, for all $i = 1, \ldots, p$. Set $d' = \max\{d(M'_i|M''_i) \mid i = 1, \ldots, p\}$. Because the M'_i's are disjoint, we have

$$d' = \sup_{x \in A^*} \min_{y \in A^*} \{d_e(x, y) \mid (x, y) \in A^*\}$$

Proceeding dually with R^{-1} we define d''. Then $k = \max\{d', d''\}$ holds. ∎

With the notations of Definition 2.4, we now show that relative distances between rational languages can be effectively computed.

Lemma 5.2 *Given two regular languages $L', L'' \subseteq A^*$, the relative distance $d_e(L'|L'')$ is effectively computable.*

Proof. If $L' \neq \emptyset$ or $L'' = \emptyset$ then $d_e(L'|L'') = \infty$ by definition. Furthermore, if $\epsilon \in L'$, then

$$d_e(L'|L'') = \max\{\min\{|u| \mid u \in L''\}, d_e(L' - \{\epsilon\}|L'')\}$$

Thus we may assume without loss of generality that $L'' \neq \emptyset$ and $\epsilon \notin L'$.

Let $\mathcal{A}' = (A, Q', \delta', q'_0, F')$ and $\mathcal{A}'' = (A, Q'', \delta'', q''_0, F'')$ be two finite deterministic automata accepting the languages L' and L'', respectively. An automaton with a weight function that computes $d(L'|L'')$ is built "on top" of the direct product of $\mathcal{A}'$ and $\mathcal{A}''$. More precisely, we have $\mathcal{A} = (A, Q, \omega, I, F)$, where $Q = Q' \times Q''$, $I = \{(q'_0, q''_0)\}$ and $F = F' \times F''$. Concerning the weight function ω, it can be viewed as encoding with each $\mathcal{A}'$-transition labeled by $a \in A$, an $\mathcal{A}''$-path of minimum distance with a. Denoting by $L''_{q'',p''}$ the set of words $u \in A^*$ labeling a path from q'' to p'' in $\mathcal{A}''$ we define

$$\omega((q', q''), a, (p', p'')) = \begin{cases} \infty & \text{if } \delta'(q', a) \neq p' \\ d_e(a, L''_{q'',p''}) & \text{otherwise.} \end{cases}$$

(Note that $d_e(a, L''_{q'',p''})$ is effectively computable.)

Fix $x = a_1a_2 \ldots a_n \in L'$ where the a_i's are letters of A. Consider $y \in L''$ such that $d_e(x, y) = d_e(x, L'')$. Applying equation (2) to the factorization $x = a_1a_2 \ldots a_n$ we have $d_e(x, y) = d_e(a_1, y_1) + \ldots + d_e(a_n, y_n)$ with $y = y_1y_2 \ldots y_n$, and $y_i \in A^*$ for $i = 1, \ldots, n$. Consider the transducer $\mathcal{T} = (A, A, Q, I, F, E)$ recognizing $L' \times L'' \subseteq A^* \times A^*$, where the transitions are of the form $((q', q''), a, b, (\delta'(q', a), \delta''(q'', b)))$ with $(a, b) \in A \times \{\epsilon\} \cup \{\epsilon\} \times A$ and $(q', q'') \in Q' \times Q''$. In the transducer $\mathcal{T}$ there exists a successful path labeled by (x, y) that can be factored as

$$(q'_0, q''_0) \xrightarrow{(a_1,y_1)} (q'_1, q''_1) \xrightarrow{(a_2,y_2)} \ldots \longrightarrow (q'_{n-1}, q''_{n-1}) \xrightarrow{(a_n,y_n)} (q'_n, q''_n) \tag{5}$$

with $q'_n \in F'$ and $q''_n \in F''$. For $i = 1, \ldots, n$ set $d_i = \omega((q'_{i-1}, q''_{i-1}), a_i, (q'_i, q''_i))$. By definition of ω we have $d_i \leq d_e(a_i, y_i)$. The path (5) yields a successful path in $\mathcal{A}$

$$(q'_0, q''_0) \xrightarrow{a_1/d_1} (q'_1, q''_1) \xrightarrow{a_2/d_2} \ldots \longrightarrow (q'_{n-1}, q''_{n-1}) \xrightarrow{a_n/d_n} (q'_n, q''_n)$$

Thus $\omega(x) \leq \sum_{i=1,\ldots,n} d_i \leq d_e(x, L'')$ holds. Conversely, consider now a successful path in $\mathcal{A}$ that yields the minimum weight

$$(q'_0, q''_0) \xrightarrow{a_1/d_1} (q'_1, q''_1) \xrightarrow{a_2/d_2} \ldots \longrightarrow (q'_{n-1}, q''_{n-1}) \xrightarrow{a_n/d_n} (q'_n, q''_n)$$

Returning to the definition of ω, choose an arbitrary y_i for $i = 1, \ldots, n$ such that $d_i = d_e(a_i, y_i)$ holds. Then there exists a path in the transducer $\mathcal{T}$

$$(q'_0, q''_0) \xrightarrow{(a_1,y_1)} (q'_1, q''_1) \xrightarrow{(a_2,y_2)} \ldots \longrightarrow (q'_{n-1}, q''_{n-1}) \xrightarrow{(a_n,y_n)} (q'_n, q''_n)$$

Then we have $d_e(x, L'') \leq \sum_{i=1,\ldots,n} d_i = \omega(x)$. This completes the proof. ∎

Previous two lemmas yield the following

Theorem 5.1 *For any recognizable relation R, it is possible to compute the smallest integer k such that R is k–reflexive with respect to the edit distance.*

Now, it is natural to investigate whether or not Theorem 5.1 can be extended to the class of synchronized rational relations. We can observe that in the case of the Hamming distance we have to compare symbols in corresponding positions. So the fact of having a synchronized transducer is very useful (see the proof of Theorem 4.1). In the case of the edit distance we have to compare symbols in different positions. Hence, it seems of no helpful to have a synchronized rational relation.

However, we are able to prove that given a constant k, the problem of deciding whether or not a synchronized rational relation is k–reflexive with respect to the edit distance is solvable. As a consequence, if we know that a synchronized rational relation R is almost reflexive, then we are able to obtain the minimum integer k such that R is k–reflexive. Note that, by Lemma 3.2, this problem is unsolvable for determinisitic rational relations.

Theorem 5.2 *For any synchronized rational relation $R \in Syn(A^* \times A^*)$ and any integer $k \geq 0$, it is possible to decide whether or not R is k–reflexive with respect to the edit distance.*

Proof. For $j \geq 0$, let $I_j \subseteq A^* \times A^*$ be the relation:

$$I_j = \{(u,v) \in A^* \times A^* \mid d_e(u,v) \leq j\}$$

Set

$$K = (A \cup \{\epsilon\}) \times (A \cup \{\epsilon\})$$

and observe that $I_0 = i_{A^*}$, $I_1 = i_{A^*} \cdot K \cdot i_{A^*}$, and $I_j = I_{j-1} \circ I_1$, where $\circ$ denotes the composition of relations. So, the relation I_1 is rational and has length difference bounded by 1, namely $||u| - |v|| \leq 1$, for $(u,v) \in I_1$. By Corollary 4.2 in [4], this implies that I_1 is a synchronized rational relation. Furthermore, the class $Syn(A^* \times A^*)$ is closed under composition. Hence, we can conclude that $I_j \in Syn(A^* \times A^*)$, for any $j \geq 0$.

Given $R \in Syn(A^* \times A^*)$ and $k \geq 0$, observe that $R \cap I_k = \{(x,y) \in R \mid d_e(x,y) \leq k\}$. Hence, R is k–reflexive if and only if the domain and the range of $R \cap I_k$ coincide with the domain and the range of R, respectively. Since the class $Syn(A^* \times A^*)$ is closed under intersection [4], and domains and ranges of rational relations are rational languages, we can conclude that it is decidable whether or not the synchronized rational relation R is k–reflexive. ∎

Open Problem Is it decidable whether or not a synchronized rational relation is almost reflexive with respect to the edit distance?

6 k–reflexivity with respect to the prefix, suffix and subword distances

In this section we extend the results of Section 4 and Section 5 to the prefix, suffix and subword distances. In fact we prove that the smallest integer k such that a recognizable relation is k–reflexive with respect to the prefix, suffix and subword distance, respectively, can be effectively computed. As in the previous section, the proofs rely on the following lemmas

Lemma 6.1 *Given two regular languages $L', L'' \subseteq A^*$, the relative distances $d_p(L'|L'')$ and $d_s(L'|L'')$ are effectively computable.*

Proof. We give the proof in the case of the prefix distance. The result for the suffix distance follows via (1). The cases $L' = \emptyset$ or $L'' = \emptyset$ are easily settled. Thus, we assume $L' \neq \emptyset$ and $L'' \neq \emptyset$ and we let $\sharp$, $\not{c}$ be two new symbols, not belonging to A.

We shall prove that there exists a weighted automaton $\mathcal{A}$ with a weight function ω whose underlying automaton recognizes the language $\sharp L' \not{c}$ and such that for any $x \in L'$, $\omega(\sharp x \not{c}) = \min\{d_p(x,y) \mid y \in L''\}$ holds. Indeed, let $\mathcal{A}' = (A \cup \{\sharp, \not{c}\}, Q', \delta', q_0', F')$ and $\mathcal{A}'' = (A \cup \{\sharp\}, Q'', \delta'', q_0'', F'')$ be two finite deterministic automata accepting the languages $\sharp L' \not{c}$ and $\sharp L''$, respectively. As in the proof of Lemma 5.2, from automata $\mathcal{A}'$ and $\mathcal{A}''$ we construct a weighted automaton $\mathcal{A}$. Each computation of $\mathcal{A}$ on input $x \in L'$, is split in two parts. In particular, in the first part, a common prefix of x and of a string belonging to L'' is recognized. To achieve this goal, the state set of $\mathcal{A}$ is the union of two isomorphic copies of the set $Q' \times Q''$.

More precisely, $\mathcal{A} = (A, Q, \omega, I, F)$, where $Q = Q' \times Q'' \times \{1,2\}$, $I = \{(q_0', q_0'', 1)\}$, $F = F' \times F'' \times \{2\}$. Now, we have to define the weight function of $\mathcal{A}$. In order to explain the definition, we consider a string $x \in L'$ and a string $y \in L''$ such that $d_p(x,y) = \min\{d_p(x,z) \mid z \in L''\}$. Then $x = \beta\gamma$, $y = \beta\mu$, where $\beta, \gamma, \mu \in A^*$, β is the longest common prefix of x and y, and $d_p(x,y) = |\gamma\mu|$.

Moreover, consider the following accepting paths of $\mathcal{A}'$ and $\mathcal{A}''$ on inputs $\sharp x \not{c}$, and $\sharp y$:

$$q_0' \xrightarrow{\sharp} r' \xrightarrow{\beta} s \xrightarrow{\gamma} s' \xrightarrow{\not{c}} p'$$

and

$$q_0'' \xrightarrow{\sharp} r'' \xrightarrow{\beta} s'' \xrightarrow{\mu} p''$$

From these paths, we want to obtain a path in $\mathcal{A}$ of the form:

$$(q_0', q_0'', 1) \xrightarrow{\sharp/0} (r, r'', 1) \xrightarrow{\beta/0} (s, s'', 2) \xrightarrow{\gamma/|\gamma|} (s', s'', 2) \xrightarrow{\not{c}/|\mu|} (p', p'', 2)$$

where in the part $(r', r'', 1) \xrightarrow{\beta/0} (s, s'', 2)$ the third component is modified in the last transition. Note that, since $d_p(x,y)$ is minimal, then μ should be a shortest string such that $p'' = \delta''(s'', \mu)$.

More precisely, the finite weights are defined as follows:

1. Part $\sharp$:
   $\omega((q_0', q_0'', 1), \sharp, (\delta'(q_0', \sharp), \delta''(q_0'', \sharp), j)) = 1$, for $j \in \{1,2\}$.

2. Part β:
   $\omega((q', q'', 1), a, (\delta'(q', a), \delta''(q'', a), j)) = 0$, for $j \in \{1,2\}$, $a \in A$.

3. Part γ:
   $\omega((q', q'', 2), a, (\delta'(q', a), q'', 2)) = 1$, for $a \in A$.

4. Part $\not{c}$:
$\omega((q',q'',2),\not{c},(p',p'',2)) = \min\{|y| \mid \delta''(q'',y)=p''\}$, for $p'=\delta'(q',\not{c})$.

The reader can easily verify that, for $q',p' \in Q'$, $q'',p'' \in Q''$, $x \in A^*$, the following equality holds:

$$\omega((q',q'',1),\sharp x \not{c},(p',p'',2)) = \begin{cases} \min\{d_p(x,y) \mid p''=\delta''(q'',y)\} & \text{if } p'=\delta'(q',x) \text{ and} \\ \infty & \text{otherwise.} \end{cases}$$

Thus, we can conclude that for $x \in L'$, $\omega(\sharp x \not{c}) = \min\{d_p(x,y) \mid y \in L''\}$ holds and $d_p(L'|L'') = \omega(\mathcal{A})$. ∎

The proof of the previous lemma can be easily adapted in order to consider the subword distance:

Lemma 6.2 *Given two regular languages $L',L'' \subseteq A^*$, the relative distance $d_f(L'|L'')$ is effectively computable.*

Proof. As in the proof of Lemma 6.1, assume $L' \neq \emptyset$ and $L'' \neq \emptyset$, and consider two symbols $\sharp$, $\not{c}$ not belonging to A.

Let $\mathcal{A}' = (A \cup \{\sharp,\not{c}\}, Q', \delta', q_0', F')$ and $\mathcal{A}'' = (A, Q'', \delta'', q_0'', F'')$ be two finite deterministic automata accepting the languages $\sharp L' \not{c}$ and L'', respectively. From automata $\mathcal{A}'$ and $\mathcal{A}''$ we construct a weighted automaton $\mathcal{A}$. Each computation of $\mathcal{A}$ on input $x \in L'$, is split in three parts, in such a way that in the second part, a common factor of x and of a string belonging to L'' is recognized. So, the state set of $\mathcal{A}$ is the union of three isomorphic copies of the set $Q' \times Q''$.

More precisely, $\mathcal{A} = (A, Q, \omega, I, F)$, where $Q = Q' \times Q'' \times \{0,1,2\}$, $I = \{(q_0', q_0'', 0)\}$, $F = F' \times F'' \times \{2\}$. To define the weight function of $\mathcal{A}$, we consider a string $x \in L'$ and a string $y \in L''$ such that $d_f(x,y) = \min\{d_f(x,z) \mid z \in L''\}$. Then $x = \alpha\beta\gamma$, $y = \eta\beta\mu$, where $\alpha,\beta,\gamma,\eta,\mu \in A^*$, β is the longest common factor of x and y, and $d_f(x,y) = |\alpha\gamma\eta\mu|$.

Moreover, consider the following accepting paths of $\mathcal{A}'$ and $\mathcal{A}''$ on inputs $\sharp x \not{c}$, and y:

$$q_0' \xrightarrow{\sharp} r' \xrightarrow{\alpha} r \xrightarrow{\beta} s \xrightarrow{\gamma} s' \xrightarrow{\not{c}} p'$$

and

$$q_0'' \xrightarrow{\eta} r'' \xrightarrow{\beta} s'' \xrightarrow{\mu} p''$$

From these paths, we want to obtain a path in $\mathcal{A}$ of the form:

$$(q_0', q_0'', 0) \xrightarrow{\sharp/|\eta|} (r', r'', 0) \xrightarrow{\alpha/|\alpha|} (r, r'', 1) \xrightarrow{\beta/0} (s, s'', 2) \xrightarrow{\gamma/|\gamma|} (s', s'', 2) \xrightarrow{\not{c}/|\mu|} (p', p'', 2)$$

where in the part $(r', r'', 0) \xrightarrow{\alpha/|\alpha|} (r, r'', 1)$ and in the part $(r, r'', 1) \xrightarrow{\beta/0} (s, s'', 2)$ the third component is modified in the last transition. Since $d_f(x,y)$ is minimal, then η (μ, resp.) should be a shortest string such that $r'' = \delta''(q_0'', \eta)$ ($p'' = \delta''(s'', \mu)$, resp.).

More precisely, the finite weights are defined as follows:

1. Part $\sharp$:
$\omega((q',q'',0),\sharp,(p',p'',j)) = \min\{|y| \mid \delta''(q'',y)=p''\}$, for $p'=\delta'(q',\sharp)$, $j \in \{0,1,2\}$.

2. Part α:
   $\omega((q', q'', 0), a, (\delta'(q', a), q'', j)) = 1$, for $j \in \{0, 1, 2\}$, $a \in A$.

3. Part β:
   $\omega((q', q'', 1), a, (\delta'(q', a), \delta''(q'', a), j)) = 0$, for $j \in \{1, 2\}$, $a \in A$.

4. Part γ:
   $\omega((q', q'', 2), a, (\delta'(q', a), q'', 2)) = 1$, for $a \in A$.

5. Part $\not{c}$:
   $\omega((q', q'', 2), \not{c}, (p', p'', 2)) = \min\{|y| \mid \delta''(q'', y) = p''\}$, for $p' = \delta'(q', \not{c})$.

The reader can easily verify that, for $q', p' \in Q'$, $q'', p'' \in Q''$, $x \in A^*$, the following equality holds:

$$\omega((q', q'', 0), \sharp x \not{c}, (p', p'', 2)) = \begin{cases} \min\{d_f(x, y) \mid p'' = \delta''(q'', y)\} & \text{if } p' = \delta'(q', x) \text{ and} \\ \infty & \text{otherwise.} \end{cases}$$

Thus, we can conclude that for $x \in L'$, $\omega(\sharp x \not{c}) = \min\{d_f(x, y) \mid y \in L''\}$ holds and $d_f(L'|L'') = \omega(\mathcal{A})$. ∎

The previous two lemmas and Lemma 5.1 yield the following

Theorem 6.1 *For any recognizable relation R, it is possible to compute the smallest integer k such that R is k–reflexive with respect to the prefix, suffix and subword distance, respectively.*

As in the case of the edit distance (Theorem 5.2), we are able to prove the following result:

Theorem 6.2 *For any synchronized rational relation $R \in Syn(A^* \times A^*)$ and any integer $k \geq 0$, it is possible to decide whether or not R is k–reflexive with respect to the prefix, suffix and subword distance, respectively.*

Proof. For $j \geq 0$, given a distance d, let $I_j \subseteq A^* \times A^*$ be the relation:

$$I_j = \{(u, v) \in A^* \times A^* \mid d(u, v) \leq j\}$$

Clearly $I_0 = i_{A^*}$. By setting $K = \{(a, \epsilon), (\epsilon, a) \mid a \in A\}$, in the case of the prefix (suffix, subword, resp.) distance we get $I_1 = i_{A^*} \cup (i_{A^*} \cdot K)$ ($I_1 = i_{A^*} \cup (K \cdot i_{A^*})$, $I_1 = i_{A^*} \cup (K \cdot i_{A^*}) \cup (i_{A^*} \cdot K)$, resp.) and more generally $I_j = I_{j-1} \circ I_1$ for all $j > 1$.

Using the same arguments of the proof of Theorem 5.2 we see that $I_j \in Syn(A^* \times A^*)$, for any $j \geq 0$, and that it is decidable, considering $R \cap I_k$, whether or not R is k–reflexive. ∎

Open Problem Is it decidable whether or not a synchronized rational relation is almost reflexive with respect to the prefix, suffix, and subword distance, respectively?

7 Unary case

In this section we restrict our attention to rational relations defined over a one letter alphabet, namely subsets of $\{a\}^* \times \{a\}^*$. We show that, unlike the general case, it is possible to compute the smallest integer k such that a relation is k–reflexive with respect to the Hamming, edit, prefix, suffix and subword distance, respectively.

For the Hamming distance the proof is quite trivial. In fact, given two integers $n, m \geq 0$, the Hamming distance between a^n and a^m is 0 when $n = m$ and is ∞ otherwise. Thus, a relation $R \subseteq \{a\}^* \times \{a\}^*$ is almost reflexive with respect to the Hamming distance if and only if it is reflexive. But this property can be easily decided [8].

We now consider the edit, prefix, suffix and subword distances. We observe that, in the unary case, these notions are equivalent. In particular, given $n, m \geq 0$,

$$d_e(a^n, a^m) = d_p(a^n, a^m) = d_s(a^n, a^m) = d_f(a^n, a^m) = |m - n|.$$

In order to state the main result of this section, we briefly recall the notion of *Presburger formula* and the correspondence between rational relations belonging to $Rat((a^*)^n)$ and Presburger formulas with n free variables over nonnegative integers [5].

Presburger formulas are built up from constants $0, 1$ and integer variables, by using the operator $+$, the predicate $=$, the logical connectives $\vee, \wedge, \neg$, and the quantifiers $\exists, \forall$. It is decidable whether or not a Presburger sentence, namely a formula without free variables, is true.

With every rational relation $R \in Rat((a^*)^n)$ it is possible to associate a formula $\phi_R(x_1, \ldots, x_n)$ with n free variables such that the n–tuples of nonnegative integers which satisfy ϕ_R exactly correspond to n–tuples of strings belonging to R, i.e., for $k_1, \ldots, k_n \in \mathbb{N}$:

$$\phi_R(k_1, \ldots, k_n) \text{ is true if and only if } (a^{k_1}, \ldots, a^{k_n}) \in R.$$

On the other hand, given a formula ψ of the Presburger arithmetic, it is possible to associate with it a rational relation R_ψ such that for $k_1, \ldots, k_n \in \mathbb{N}$:

$$(a^{k_1}, \ldots, a^{k_n}) \in R_\psi \text{ if and only if } \psi(k_1, \ldots, k_n) \text{ is true.}$$

At this point, we are able to prove the main result of this section:

Theorem 7.1 *For any rational relation R defined over a one letter alphabet, it is possible to compute the smallest integer k such that R is k–reflexive with respect to the edit, prefix, suffix and subword distance, respectively.*

Proof. Given $R \in Rat(\{a\}^* \times \{a\}^*)$, let $S' \subseteq \mathbb{N}$ be the set containing exactly all integers which represent minimum distances between a string $a^n \in Dom(R)$ and a string a^p such that $(a^n, a^p) \in R$, namely:

$$S' = \{\min \{|n - p| \mid (a^n, a^p) \in R\} \mid a^n \in Dom(R)\}.$$

In a similar way, we define

$$S'' = \{\min \{|m - p| \mid (a^p, a^m) \in R\} \mid a^m \in Ran(R)\}.$$

It is not difficult to see that, given an integer k, R is k–reflexive if and only if both $\sup(S')$ and $\sup(S'')$ are bounded by k.

Now, we observe that S' can be rewritten as:

$$\begin{aligned} S' &= \{|n-m| \mid (a^n, a^m) \in R \wedge \forall h((a^n, a^p) \in R \to (|n-m| \le |n-p|))\} \\ &= \{x \mid \exists n \exists m((x = |n-m|) \wedge (a^n, a^m) \in R \\ &\qquad \wedge \forall p \exists y(((|n-p| = y) \wedge (a^n, a^p) \in R) \to (x \le y)))\}. \end{aligned}$$

Hence, S' is the set of integers which satisfy the following Presburger formula:

$$\rho'(x) = \exists n \exists m(abs(n,m,x) \wedge \phi_R(n,m) \wedge \forall p \exists y((abs(n,p,y) \wedge \phi_R(n,p)) \to (x \le y))),$$

where ϕ_R is the formula associated with the relation R. The abbreviation $abs(x,y,z)$ stands for $z = |x-y|$ and can be translated to

$$abs(x,y,z) = ((x < y) \to (x+z=y)) \wedge ((x \ge y) \to (y+z=x)).$$

where $x \le y$ stands for $\exists z(y = x+z)$. Thus, the (unary) relation $R_{\rho'}$ associated with ρ' is a rational language. This implies that $\sup(S')$ and, in a similar way, $\sup(S'')$ are computable. ∎

References

[1] J. Berstel, *Transductions and context–free languages.* Teubner, Stuttgart, 1979.

[2] S. Eilenberg. *Automata, languages and machines*, volume A. Academic Press, 1974.

[3] C. Elgot and J. Mezei. On Relations Defined by Generalized Finite Automata. *IBM J. Res. Develop.*, 9:47–68, 1965.

[4] C. Frougny and J. Sakarovitch. Synchronized rational relations of finite and infinite words. *Theoretical Computer Science*, 108:45–82, 1993.

[5] S. Ginsburg and E. Spanier. Semigroups, Presburger formulas, and languages. *Pacific Journal of Mathematics*, 16:285–296, 1966.

[6] K. Hashiguchi. Limitedness theorem on finite automata with distance functions. *Journal of Computer and System Sciences*, 24:233–244, 1982.

[7] K. Hashiguchi. Improved limitedness theorems on finite automata with distance functions. *Theoretical Computer Science*, 72:27–38, 1990.

[8] J. Johnson. Rational equivalence relations. *Theoretical Computer Science*, 47:39–60, 1986.

[9] J. Kruskal. An overview of sequence comparison. In D. Sankoff and J. Kruskal, editors, *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, pages 1–44. Addison–Wesley, 1983.

[10] M. Rabin and D. Scott. Finite automata and their decision problems. *IBM J. Res. Develop*, 3:125–144, 1959. Also in: E.F. Moore. *Sequential machines.* Addison–Wesley, 1964.

[11] K. Ruohonen. Reversible machines and Post's correspondence problem for biprefix morphisms. *J. Inform. Process. Cybernet. EIK*, 21:579–595, 1985.