Linear Systems for Regular Hedge Languages

Mircea Marin[1]* and Temur Kutsia[2]**

[1] Department of Computer Science, University of Tsukuba, Japan
[2] RISC, Johannes Kepler University, Linz, Austria

Abstract. We propose linear systems of hedge language equations (LSH) as a formalism to represent regular hedge languages. These linear systems are suitable for several computations in the algebra of regular hedge languages. We indicate algorithms to translate between representations by hedge automata and LSH, and for the computation of LSH for the intersection, quotient, left and right factors of regular hedge languages.

1 Introduction

Regular hedge languages (RHLs) play an important role in computer science where they are well known as a formalism for a schema of XML [8]. There are many equivalent ways to represent RHLs: by hedge automata [7], regular hedge grammars [6], regular hedge expressions [7], regular expression types for XML [3], etc. The choice of a suitable representation depends on the computation under consideration, and conversions between representations are often required.

We propose a new characterization of RHLs, by linear systems of hedge language equations (LSH for short). LSHs can be viewed as a generalization of the notion of system of linear equations over a Kleene algebra [4] which is linear in both horizontal and vertical directions. An important result is that LSHs have a unique solution and that the solution consists of regular hedge languages. Solving LSHs can be achieved by a slight generalization of solving linear systems over a Kleene algebra. Conversely, for every language L represented by a hedge automaton we can compute an LSH with variables $x_1, \ldots, x_n$ whose solution for x_1 coincides with L. Thus, we can use LSHs to represent RHLs.

LSHs are convenient for several computations in the algebra of RHLs. Many properties of regular word languages carry over to RHLs, such as closure under intersection and quotient, and the fact that the factors of RHLs are regular and finitely many. In this paper we indicate how LSHs can be used to compute the intersection, quotient, left and right factors of regular hedge languages.

The paper is structured as follows. In Sect. 2 we define LSHs and provide algorithms to translate between LSH and hedge automaton. Sections 3–5 describe algorithms for the computation of intersection, right quotient, and left factors of RHLs represented by LSHs. Section 6 concludes.

* Supported by JSPS Grant-in-Aid no. 20500025 for Scientific Research (C).

** Supported by EC FP6 under the project SCIEnce—Symbolic Computation Infrastructure for Europe (Contract No. 026133).

2 Linear Systems of Hedge Language Equations

For any set S, by 2^S we mean the set of all subsets of S. For any finite set A we consider the set A^* of all finite words over A, and denote the empty word by ϵ. The set $\mathbf{Reg}(A)$ of regular expressions over A is defined by the grammar $r ::= 0 \mid 1 \mid a \mid r + r \mid r\,r \mid r^\star$ where $a \in A$. We write $[\![r]\!]$ for the usual interpretation of $r \in \mathbf{Reg}(A)$ as a regular language, and $r_1 \doteq r_2$ if $[\![r_1]\!] = [\![r_2]\!]$. The *constant part* $\mathsf{o}(r)$ of r is defined recursively on the structure of r such that it is 1 if $\epsilon \in [\![r]\!]$ and 0 otherwise [1].

Hedges over an alphabet Σ with constants from a set $\mathcal{K}$ are finite sequences of trees produced by the grammar $h ::= \epsilon \mid k \mid a\langle h\rangle\, h$ where $a \in \Sigma$ and $k \in \mathcal{K}$. We denote this set by $\mathcal{H}(\Sigma, \mathcal{K})$. A *hedge language* (HL) is a set of hedges. The product of two HLs L and M is the HL $L\,M := \{h\,h' \mid h \in L, h' \in M\}$.

In this paper we consider only HLs with no constants. We also consider an infinite set $\mathcal{X}$ of hedge language variables and regular hedge expressions over Σ and $\mathcal{X}$ generated by $w ::= 0 \mid 1 \mid x \mid a\langle w\rangle \mid w + w \mid w\,w \mid w^\star$ where $a \in \Sigma$ and $x \in \mathcal{X}$. An *assignment* is a mapping σ from variables to HLs. Given an assignment σ, we interpret regular hedge expressions over Σ and $\mathcal{X}$ as follows: $[\![0]\!]_\sigma := \emptyset$, $[\![1]\!]_\sigma := \{\epsilon\}$, $[\![x]\!]_\sigma := \sigma(x)$, $[\![w_1 + w_2]\!]_\sigma := [\![w_1]\!]_\sigma \cup [\![w_2]\!]_\sigma$, $[\![w_1\,w_2]\!]_\sigma := [\![w_1]\!]_\sigma\,[\![w_2]\!]_\sigma$, $[\![a\langle w\rangle]\!]_\sigma := \{a\langle h\rangle \mid h \in [\![w]\!]_\sigma\}$, and $[\![w^\star]\!]_\sigma := \bigcup_{n=0}^{\infty} [\![w]\!]_\sigma^n$ where $[\![w]\!]_\sigma^0 := \{\epsilon\}$ and $[\![w]\!]_\sigma^n := \{h_1 \ldots h_n \mid h_1, \ldots, h_n \in [\![w]\!]_\sigma\}$ for $n \geq 1$. Also, we write $w_1 \doteq_\sigma w_2$ if $[\![w_1]\!]_\sigma = [\![w_2]\!]_\sigma$.

A *hedge automaton* (HA) is a 4-tuple $\mathcal{A} = (\Sigma, \mathcal{Q}, \mathcal{P}, r_1)$ where Σ is the alphabet for hedges, $\mathcal{Q}$ is a finite set of states, $r_1 \in \mathbf{Reg}(\mathcal{Q})$, and $\mathcal{P}$ is a finite set of transition rules of the form $\mathsf{q} \to a\langle r\rangle$ with $\mathsf{q} \in \mathcal{Q}$, $a \in \Sigma$, and $r \in \mathbf{Reg}(\mathcal{Q})$. The language *accepted* by $\mathcal{A}$ is the set $L(\mathcal{A}) := \{h \in \mathcal{H}(\Sigma, \emptyset) \mid h \to_{\mathcal{P}}^* v \wedge v \in [\![r_1]\!]\}$, where $\to_{\mathcal{P}}$ is the transition relation induced by $\mathcal{P}$ on $\mathcal{H}(\Sigma, \mathcal{Q})$. A hedge language is *regular* (*RHL* for short) if it is accepted by a hedge automaton.

A *linear system of hedge language equations* (LSH) over a finite alphabet Σ with variables from $\{x_1, \ldots, x_n\}$ is a system of equations of the form

$$x_i = \ell_{i1}\,x_1 + \ldots + \ell_{in}\,x_n + b_i \qquad (1 \leq i \leq n) \tag{1}$$

with ℓ_{ij} sums of elements from $\{a\langle x_l\rangle \mid a \in \Sigma, 1 \leq l \leq n\}$ and $b_i \in \{0, 1\}$ for all $i, j \in \{1, \ldots, n\}$. If $\ell_{ij} \neq 0$ then we say that x_j occurs at *horizontal position* in the right side of the equation of x_i. A *solution* of (1) is an assignment σ for $\mathcal{X} = \{x_1, \ldots, x_n\}$ such that $x_i \doteq_\sigma \ell_{i1}\,x_1 + \ldots + \ell_{in}\,x_n + b_i$ for all $1 \leq i \leq n$.

Solving Linear Systems of Hedge Language Equations. Suppose $\Sigma = \{a_1, \ldots, a_p\}$ and σ is a solution of (1). We solve (1) in two steps:

Abstraction step. Let $\mathcal{Q} := \{\mathsf{q}_{kl} \mid 1 \leq k \leq p, 1 \leq l \leq n\}$ be a set of fresh symbols. We replace every coefficient $a_i\langle x_j\rangle$ of (1) with q_{ij}. This replacement produces a linear system of equations over the Kleene algebra $\mathbf{Reg}(\mathcal{Q})$:

$$x_i = m_{i1}\,x_1 + \ldots + m_{in}\,x_n + b_i \qquad (1 \leq i \leq n)$$

where m_{ij} are sums of elements from $\mathcal{Q}$, and $b_i \in \{0, 1\}$.

Solving step. Let $\mathcal{P} := \{a_k\langle r_l\rangle \to \mathtt{q}_{kl} \mid 1 \leq k \leq p, 1 \leq l \leq n\}$, and compute

$$\begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} := M^\star \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \quad \text{where } M = \begin{pmatrix} m_{11} & \dots & m_{1n} \\ \vdots & \ddots & \vdots \\ m_{n1} & \dots & m_{nn} \end{pmatrix}$$

and $M^\star$ is the asterate of matrix M [4].
The unique solution of (1) is $\{x_1 \mapsto L_1, \dots, x_n \mapsto L_n\}$ where, for every $1 \leq i \leq n$, L_i is the language accepted by the HA $(\Sigma, \mathcal{Q}, \mathcal{P}, r_i)$.

The correctness of this algorithm can be explained as follows. Let μ be the extension of σ to $\mathcal{X} \cup \mathcal{Q}$ with the assignments $\mu(\mathtt{q}_{kl}) := [\![a_k\langle x_l\rangle]\!]_\sigma$ for all $\mathtt{q}_{kl} \in \mathcal{Q}$. Then $x_i \doteq_\mu m_{i1}\,x_1 + \ldots + m_{in}\,x_n + b_i$ for $1 \leq i \leq n$. Since $[\![m_{ij}]\!]_\mu \subseteq \bigcup_{k=1}^{p} \bigcup_{l=1}^{n} [\![a_k\langle x_j\rangle]\!]_\sigma$ for all i, j, we learn that m_{ij} denote languages of terms, which are ϵ-free HLs. By [5, Lemma 1], we have $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \doteq_\mu M^\star \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$, thus $x_i \doteq_\mu r_i$ for $1 \leq i \leq n$. This relation shows that the solution of (1) is unique and that, for every $1 \leq i \leq n$, $[\![x_i]\!]_\sigma$ coincides with the language recognized by the HA $(\Sigma, \mathcal{Q}, \mathcal{P}, r_i)$ where $\mathcal{P} = \{a_k\langle r_l\rangle \to \mathtt{q}_{kl} \mid 1 \leq k \leq p, 1 \leq l \leq n\}$.

Example 1. The equations $x_1 = (a_1\langle x_1\rangle + a_2\langle x_2\rangle)\,x_1 + a_1\langle x_1\rangle\,x_2$ and $x_2 = a_2\langle x_2\rangle\,x_2 + 1$ form an LSH over signature $\Sigma = \{a_1, a_2\}$ can be solved as follows. First we abstract the coefficients $a_1\langle x_1\rangle$ and $a_2\langle x_2\rangle$ by replacing them with $\mathtt{q}_{11}$ and $\mathtt{q}_{22}$ respectively. This replacement produces the new system of equations $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = M \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ where $M = \begin{pmatrix} \mathtt{q}_{11} + \mathtt{q}_{22} & \mathtt{q}_{11} \\ 0 & \mathtt{q}_{22} \end{pmatrix}$. Then

$$M^\star \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} (\mathtt{q}_{11} + \mathtt{q}_{22})^\star & (\mathtt{q}_{11} + \mathtt{q}_{22})^\star \mathtt{q}_{11} \mathtt{q}_{22}^\star \\ 0 & \mathtt{q}_{22}^\star \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} (\mathtt{q}_{11} + \mathtt{q}_{22})^\star \mathtt{q}_{11} \mathtt{q}_{22}^\star \\ \mathtt{q}_{22}^\star \end{pmatrix}$$

and we conclude that the solution of this LSH is the assignment σ such that $\sigma(x_1) = L((\Sigma, \{\mathtt{q}_{11}, \mathtt{q}_{22}\}, \mathcal{P}, r_1))$ and $\sigma(x_2) = L((\Sigma, \{\mathtt{q}_{11}, \mathtt{q}_{22}\}, \mathcal{P}, r_2))$, where $r_1 := (\mathtt{q}_{11} + \mathtt{q}_{22})^\star \mathtt{q}_{11} \mathtt{q}_{22}^\star$, $r_2 := \mathtt{q}_{22}^\star$, and $\mathcal{P} := \{a_1\langle r_1\rangle \to \mathtt{q}_{11}, a_2\langle r_2\rangle \to \mathtt{q}_{22}\}$. □

Since an LSH has unique solution, we can define the notion of *LSH for a hedge language* L as any LSH whose solution σ assigns language L to the variable that occurs in the left hand side of its first equation.

Converting HA into LSH. Suppose $\mathcal{A} = (\Sigma, \mathcal{Q}, \mathcal{P}, r_1)$ is an HA and $\Sigma = \{a_1, \dots, a_p\}$. We indicate how to compute an LSH over Σ and a set of variables $\{x_1, \dots, x_n\}$ such that its unique solution σ has $\sigma(x_1) = L(\mathcal{A})$.

Let $R := \{r_1\} \cup \{r \mid \exists a\langle r\rangle \to \mathtt{q} \in \mathcal{P}\}$. It is well known [1] that for any regular expression $r \in \mathbf{Reg}(\mathcal{Q})$ we can compute: (1) a finite set $\partial_{\mathcal{Q}^*}(r)$ of regular expressions in $\mathbf{Reg}(\mathcal{Q}) \setminus \{0\}$, and (2) for every $s \in \partial_{\mathcal{Q}^*}(r)$, a finite set $\mathtt{lf}(s)$

of pairs $\langle \mathtt{q}, s' \rangle \in \mathcal{Q} \times \partial_{\mathcal{Q}^*}(r)$, such that $s \doteq \mathtt{o}(s) + \sum_{\langle \mathtt{q}, s' \rangle \in \mathtt{lf}(s)} \mathtt{q}\, s'$. Thus, if $\{r_1, \ldots, r_n\} := \bigcup_{r \in R} \partial_{\mathcal{Q}^*}(r)$ then $r_i \doteq \mathtt{o}(r_i) + \sum_{\langle \mathtt{q}, r \rangle \in \mathtt{lf}(r_i)} \mathtt{q}\, r$ for $1 \leq i \leq n$.

Let $\mathcal{X} := \{x_i \mid 1 \leq 1 \leq n\}$ be a set of fresh variables, and the assignment σ of variables from $\mathcal{X} \cup \mathcal{Q}$ such that $\sigma(x_i)$ is the language accepted by $(\Sigma, \mathcal{Q}, \mathcal{P}, r_i)$ for all $1 \leq i \leq n$, and $\sigma(\mathtt{q})$ is the language accepted by $(\Sigma, \mathcal{Q}, \mathcal{P}, \mathtt{q})$, for all $\mathtt{q} \in \mathcal{Q}$. If we replace every horizontal occurrence of r_i with x_i in the previous equations, we obtain $x_i \doteq_\sigma b_i + \sum_{j=1}^{n} m_{ij}\, x_j$ for $1 \leq i \leq n$, where $b_i = \mathtt{o}(r_i) \in \{0, 1\}$ and m_{ij} are sums of elements of $\mathcal{Q}$ for all $1 \leq i, j \leq n$. We define the regular hedge expressions $\mathtt{re}(\mathtt{q}) := \sum_{(a\langle r_i \rangle \to \mathtt{q}) \in \mathcal{P}} a\langle x_i \rangle$ for all $\mathtt{q} \in \mathcal{Q}$, and $\mathtt{re}(m_{ij}) := \sum_{\mathtt{q} \in S_{ij}} \mathtt{re}(\mathtt{q})$ where S_{ij} is the subset of $\mathcal{Q}$ for which $m_{ij} = \sum_{\mathtt{q} \in S_{ij}} \mathtt{q}$. Then obviously $m_{ij} \doteq_\sigma \mathtt{re}(m_{ij})$ for all $i, j \in \{1, \ldots, n\}$, and thus we have $x_i \doteq_\sigma b_i + \sum_{j=1}^{n} \mathtt{re}(m_{1j})\, x_j$ for $1 \leq i \leq n$. Since $\mathtt{re}(\mathtt{m}_{ij})$ are sums of regular hedge expressions from $\{a\langle x \rangle \mid a \in \Sigma, x \in \mathcal{X}\}$, what we've got is an LSH over Σ with variables $x_1, \ldots, x_n$ whose unique solution is the restriction of σ to $\mathcal{X}$. The first component of the solution of this LSH is $\sigma(x_1)$, which is $L(\mathcal{A})$.

Example 2. Consider the HA $\mathcal{A} = (\Sigma, \{\mathtt{q}_{11}, q_{22}\}, \mathcal{P}, (\mathtt{q}_{11} + \mathtt{q}_{22})^\star \mathtt{q}_{11} \mathtt{q}_{22}^\star)$ where $\Sigma = \{a_1, a_2\}$ and $\mathcal{P} = \{a_1\langle (\mathtt{q}_{11} + \mathtt{q}_{22})^\star \mathtt{q}_{11} \mathtt{q}_{22}^\star \rangle \to \mathtt{q}_{11}, a_2\langle \mathtt{q}_{22}^\star \rangle \to \mathtt{q}_{22}\}$.

This is the HA computed in Example 1 from an LSH with 2 equations. In this example we have $R = \{r_1, r_2\}$ where $r_1 := (\mathtt{q}_{11} + \mathtt{q}_{22})^\star \mathtt{q}_{11} \mathtt{q}_{22}^\star$ and $r_2 := \mathtt{q}_{22}^\star$, and $\partial_{\mathcal{Q}^*}(r_1) \cup \partial_{\mathcal{Q}^*}(r_2) = R$. We have $\mathtt{o}(r_1) = 0$, $\mathtt{o}(r_2) = 1$, $r_1 \doteq (\mathtt{q}_{11} + \mathtt{q}_{22})\, r_1 + \mathtt{q}_{11}\, r_2$, $r_2 \doteq \mathtt{q}_{22}\, r_2 + 1$, and $\mathtt{re}(\mathtt{q}_{11} + \mathtt{q}_{22}) = \mathtt{re}(\mathtt{q}_{11}) + \mathtt{re}(\mathtt{q}_{22}) = a_1\langle x_1 \rangle + a_2\langle x_2 \rangle$, $\mathtt{re}(\mathtt{q}_{11}) = a_1\langle x_1 \rangle$, $\mathtt{re}(\mathtt{q}_{22}) = a_2\langle x_2 \rangle$. We obtain the equations $x_1 = (a_1\langle x_1 \rangle + a_2\langle x_2 \rangle)\, x_1 + a_1\langle x_1 \rangle\, x_2$ and $x_2 = a_2\langle x_2 \rangle\, x_2 + 1$ which form an LSH whose unique solution σ satisfies the condition $\sigma(x_1) =$ the language of $\mathcal{A}$. □

3 Intersection of Regular Hedge Languages

In this section we indicate how to compute an LSH for $L \cap M$ from LSHs for L and M. Let's assume given an LSH S made of equations $x_i = c_i + \sum_{k=1}^{m} a_{ik}\, x_k$ $(1 \leq i \leq m)$ and with solution σ such that $\sigma(x_1) = L$, and an LSH T made of equations $y_j = d_j + \sum_{l=1}^{n} b_{jl}\, y_l$ $(1 \leq j \leq n)$ with solution τ such that $\tau(y_1) = M$, and that $c_i, d_j \in \{0, 1\}$, a_{ik} are sums of elements from $\{a\langle x_u \rangle \mid a \in \Sigma, 1 \leq u \leq m\}$, and b_{jl} are sums of elements from $\{a\langle y_v \rangle \mid a \in \Sigma, 1 \leq v \leq n\}$. The idea of computing an LSH for $L \cap M$ is based on the principle of intersecting equations of S with equations of T. When we intersect $x_i = c_i + \sum_{k=1}^{m} a_{ik}\, x_k$ with $y_j = d_j + \sum_{l=1}^{n} b_{jl}\, y_l$, we aim at computing an equation that characterizes the intersection of RHLs $\sigma(x_i) \cap \mu(y_j)$. We regard the set of expressions $\mathcal{Z} := \{x_k \cap y_l \mid 1 \leq k \leq m, 1 \leq l \leq n\}$ as variables and consider the assignment ν for variables from $\mathcal{Z}$ defined by $\nu(x_k \cap y_l) := \sigma(x_k) \cap \mu(y_l)$ for all $1 \leq k \leq m$ and $1 \leq l \leq n$. Since $x_i \doteq_\sigma c_i + \sum_{k=1}^{m} a_{ik}\, x_k$ and $y_j \doteq_\mu d_j + \sum_{l=1}^{n} b_{jl}\, y_l$, we can compute regular hedge expressions s_{ijkl} such that $x_i \cap y_j \doteq_\nu \min(c_i, d_j) + \sum_{k=1}^{m} \sum_{l=1}^{n} s_{ijkl}\, (x_k \cap y_l)$ for $1 \leq i \leq m$ and $1 \leq j \leq n$, where s_{ijkl} are sums of regular hedge expressions of the form $a\langle z \rangle$ with $a \in \Sigma$ and $z \in \mathcal{Z}$. More precisely:

- We identify two families of finite sets $\{U_{ik} \mid 1 \leq i,k \leq m\} \in 2^{\Sigma \times \{x_1,\ldots,x_m\}}$ and $\{V_{jl} \mid 1 \leq j,l \leq n\} \in 2^{\Sigma \times \{y_1,\ldots,y_n\}}$ such that $a_{ik} = \sum_{\langle a,u\rangle \in U_{ik}} a\langle x_u\rangle$ and $b_{jl} = \sum_{\langle a,v\rangle \in V_{jl}} a\langle y_v\rangle$ for all $1 \leq i,k \leq m$ and $1 \leq j,l \leq n$.
- We define $s_{ijkl} := \sum_{a \in \Sigma} \sum_{\langle a,x_u\rangle \in U_{ik} \wedge \langle a,y_v\rangle \in V_{jl}} a\langle x_u \cap y_v\rangle$.

For example, the intersection of the equations $x_1 = 1 + (a\langle x_1\rangle + b\langle x_3\rangle)\, x_1 + (b\langle x_3\rangle + d\langle x_4\rangle)\, x_2$ and $y_2 = 1 + (a\langle y_1\rangle + c\langle y_2\rangle)\, y_1 + b\langle y_4\rangle\, y_2$ produces the equation $x_1 \cap y_2 = 1 + a\langle x_1 \cap y_1\rangle\, (x_1 \cap y_1) + b\langle x_3 \cap y_4\rangle\, (x_1 \cap y_2) + b\langle x_3 \cap y_4\rangle\, (x_2 \cap y_2)$.

We can construct an LSH I for the HL $L \cap M = [\![x_1 \cap y_1]\!]_\nu$ as follows:

1. Intersect the first equation of S with the first equation of T and add it to I. This intersection produces an equation with variable $x_1 \cap y_1$ to the left.
2. For every variable $x_k \cap y_l$ that occurs in the right side of some equation already in I, add to I the intersection of the equation for x_k in S with the equation for y_l in T.

This process will terminate because $\mathcal{Z}$ is a finite set, so we can not add indefinitely equations to I. We end up with an LSH of at most $m \times n$ equations for the RHL $[\![x_1 \cap y_1]\!]_\nu = L \cap M$.

4 Quotient of Regular Hedge Languages

The *quotient* of an HL L with respect to an HL M is the HL $M^{-1}L := \{h \mid \exists h' \in M$ such that $h'\, h \in L\}$. Like for regular languages, we can prove that if L is RHL and M is *any* HL then $M^{-1}L$ is RHL. To see why this is so, assume L is the language recognized by an HA $(\Sigma, \mathcal{Q}, \mathcal{P}, r)$ and let $\{r_1, \ldots, r_n\} := \bigcup_{w \in \mathcal{Q}^*} \partial_w(r)$. It can be shown for any hedge h, the HL $\{h\}^{-1}L$ is recognized by an HA from $\left\{(\Sigma, \mathcal{Q}, \mathcal{P}, \sum_{s \in Q'} s) \;\middle|\; Q' \subseteq \{r_1, \ldots, r_n\}\right\}$. This is a finite set of at most $2^{\|r\|+1}$ HAs, where $\|r\|$ is the alphabetic width of $r \in \mathbf{Reg}(\mathcal{Q})$ [1, Corollary 10]. Thus, $\{\{h\}^{-1}L \mid h \in M\}$ is a finite set of RHLs. But $M^{-1}L = \bigcup_{h \in M} \{h\}^{-1}L$ is a finite union of RHLs, hence it is RHL too.

Similarly, we can define the *right quotient* of an HL L with respect to an HL M as the HL $LM^{-1} := \{h \mid \exists h' \in M$ such that $h\, h' \in L\}$. If we define the *symmetric* L^{s} of L as the language obtained by reversing the order of trees at the outermost level in hedges, then $(L^{\mathsf{s}})^{\mathsf{s}} = L$ and $M^{-1}L = (L^{\mathsf{s}}(M^{\mathsf{s}})^{-1})^{\mathsf{s}}$ for any HLs L and M. Moreover, if L is RHL then L^{s} is RHL too. Since $M^{-1}L = (L^{\mathsf{s}}(M^{\mathsf{s}})^{-1})^{\mathsf{s}}$, we can achieve quotient computations via right quotient computations. Therefore, in the remainder of this section we consider only the computation of right quotient.

If M is RHL then we can compute a representation of LM^{-1}. In the remainder of this section we indicate a method to compute an LSH for LM^{-1} when we know an LSH for L and an LSH for M.

Suppose the LSHs for L and M are like in the previous section, $\mathcal{X} := \{x_1, \ldots, x_m\}$, $\mathcal{Y} := \{y_1, \ldots, y_n\}$, and let σ and μ be their unique solutions. Let $\mathcal{Z} := \{z_1, \ldots, z_n\}$ be a set of fresh new variables and ν the assignment for

$\mathcal{X} \cup \mathcal{Z}$ which extends σ by associating every z_i with the language $\llbracket x_i \rrbracket_\sigma \llbracket y_1 \rrbracket_\mu^{-1}$. We can construct incrementally an LSH S for LM^{-1} as follows:

1. Since $x_1 \doteq_\sigma c_1 + \sum_{k=1}^{m} a_{1k}\, x_k$, we can multiply it to the right with $\llbracket y_1 \rrbracket_\mu^{-1}$ and obtain $z_1 \doteq_\nu e_1 + \sum_{k=1}^{m} a_{1k}\, z_k$ where $e_1 = 1$ if $\epsilon \in \llbracket x_1 \rrbracket_\sigma \llbracket y_1 \rrbracket_\mu^{-1}$ and $e_1 = 0$ otherwise. We add to S the equation $z_1 = e_1 + \sum_{k=1}^{m} a_{1k}\, z_k$.
2. For every variable $z_i \in \mathcal{Z}$ which occurs in the right side of some equation already in S, add to S the equation $z_i = e_i + \sum_{k=1}^{m} a_{ik}\, z_k$ with $e_i = 1$ if $\epsilon \in \llbracket x_i \rrbracket_\sigma \llbracket y_1 \rrbracket_\mu^{-1}$ and $e_i = 0$ otherwise. The addition of this equation to S is justified by the relation $z_i \doteq_\nu e_i + \sum_{k=1}^{m} a_{ik}\, z_k$ whose validity follows by multiplying the relation $x_i \doteq_\sigma c_i + \sum_{k=1}^{m} a_{ik}\, x_k$ to the right with $\llbracket y_1 \rrbracket_\mu^{-1}$. This process will eventually terminate because $\mathcal{Z}$ is a finite set, so we do not add indefinitely equations to S.
3. Finally, we add to S the equations of the LSH for L.

We end up with S being an LSH containing equations $z_{i_j} = e_{i_j} + \sum_{k \in I} a_{i_j k}\, z_k$ $(1 \le j \le p)$ in addition to the equations of the LSH for L, where $I = \{i_1, \dots, i_p\} \subseteq \{1, \dots, m\}$ with $i_1 = 1$. This is an LSH with at most $2\,m$ equations whose unique solution of S is the restriction of ν to $\{z_i \mid i \in I\} \cup \mathcal{X}$. To compute $e_{i_1}, \dots, e_{i_p}$ it is useful to notice that for every $1 \le i \le m$ we have $e_i = 1$ iff $\epsilon \in \llbracket z_i \rrbracket_\nu = \llbracket x_i \rrbracket_\sigma \llbracket y_1 \rrbracket_\mu^{-1}$ iff $\llbracket x_i \rrbracket_\sigma \cap \llbracket y_1 \rrbracket_\mu \neq \emptyset$. Since $\llbracket x_i \rrbracket_\sigma = \llbracket c_i \rrbracket \cup \bigcup_{k=1}^{m} \llbracket a_{ik} \rrbracket_\sigma \llbracket x_k \rrbracket_\sigma$ and $\llbracket y_j \rrbracket_\mu = \llbracket d_j \rrbracket \cup \bigcup_{l=1}^{n} \llbracket b_{jl} \rrbracket_\mu \llbracket y_l \rrbracket_\mu$, we learn that $\llbracket x_i \rrbracket_\sigma \cap \llbracket y_j \rrbracket_\mu \neq \emptyset$ iff

1. $c_i = d_j = 1$ (in this case, $\epsilon \in \llbracket x_i \rrbracket_\sigma \cap \llbracket y_j \rrbracket_\mu$), or
2. there exist $k \in \{1, \dots, m\}$ and $l \in \{1, \dots, n\}$ such that $\llbracket a_{ik} \rrbracket_\sigma \cap \llbracket b_{jl} \rrbracket_\mu \neq \emptyset$ and $\llbracket x_k \rrbracket_\sigma \cap \llbracket y_l \rrbracket_\mu \neq \emptyset$.

It follows that $\llbracket x_u \rrbracket_\sigma \cap \llbracket y_v \rrbracket_\mu \neq \emptyset$ iff the judgment $x_u \diamond y_v$ can be inferred with

$$\frac{[c_i = 1 \wedge d_j = 1]}{x_i \diamond y_j} \qquad \frac{x_{i_1} \diamond y_{j_1} \quad x_{i_2} \diamond y_{j_2} \quad [a\langle x_{i_1} \rangle\, x_{i_2} \in \mathrm{rhs}(x_i) \wedge a\langle y_{j_1} \rangle y_{j_2} \in \mathrm{rhs}(y_j)]}{x_i \diamond y_j}$$

where $a\langle x_{i_1} \rangle\, x_{i_2} \in \mathrm{rhs}(x_i)$ means that $a\langle x_{i_1} \rangle\, x_{i_2}$ occurs in the right side of the equation for x_i, and the meaning of $a\langle y_{j_1} \rangle\, y_{j_2} \in \mathrm{rhs}(y_i)$ is that $a\langle y_{j_1} \rangle\, y_{j_2}$ occurs as summand in the right hand side of the equation for y_j. In particular $e_i = 1$ iff the judgment $x_i \diamond y_1$ can be inferred with the inference rules mentioned above.

Example 3. Consider the LSHs

$$\begin{array}{ll}
x_1 = 1 + a_1\langle x_2 \rangle\, x_1 + a_2\langle x_4 \rangle\, x_2 & \qquad y_1 = 1 + (a_1\langle y_1 \rangle + a_2\langle y_2 \rangle + a_3\langle y_1 \rangle)\, y_1 \\
x_2 = 1 + a_2\langle x_4 \rangle\, x_2 & \qquad y_2 = (a_1\langle y_1 \rangle + a_2\langle y_2 \rangle + a_3\langle y_1 \rangle)\, y_3 \\
x_3 = (a_2\langle x_4 \rangle + a_3\langle x_4 \rangle)\, x_4 & \qquad y_3 = 1 \\
x_4 = 1 + a_1\langle x_4 \rangle\, x_4 &
\end{array}$$

with solutions σ and μ. Our construction of an LSH for $\llbracket x_1 \rrbracket_\sigma \llbracket y_1 \rrbracket_\mu^{-1}$ yields the LSH with the equations $z_1 = e_1 + a_1\langle x_2 \rangle\, z_1 + a_2\langle x_4 \rangle\, z_2$ and $z_2 = e_2 + a_2\langle x_4 \rangle\, z_2$ besides the equations of the first LSH, where e_1 and e_2 are still to be computed. In this example, $e_1 = e_2 = 1$ because the inference rules

$$\frac{}{x_1 \diamond y_1} \qquad \frac{x_4 \diamond y_2 \quad x_2 \diamond y_3}{x_1 \diamond y_2} \qquad \frac{x_4 \diamond y_1 \quad x_4 \diamond y_3}{x_4 \diamond y_2} \qquad \frac{}{x_2 \diamond y_3} \qquad \frac{}{x_4 \diamond y_1} \qquad \frac{}{x_4 \diamond y_3}$$

are available to infer the judgments $x_1 \diamond y_1$ and $x_1 \diamond y_2$. □

Note that in this example we had $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$, but we had to add only two equations, those for z_1 and z_2, in the LSH for $[\![x_1]\!]_\sigma [\![y_1]\!]_\mu^{-1}$. In general, we can say that the number of equations that must be prepended to the LSH for L to produce an LSH for LM^{-1} is between 1 and the number of variables of $\mathcal{X}$ that occur at horizontal positions in the equations of the LSH for L.

5 Left and Right Factors of a Regular Hedge Language

The following are straightforward generalizations to HLs of notions from Conway's theory of factorizations of regular languages [2]. A product of HLs $F_1 \ldots F_n$ is a *subfactorization* of a an HL E if and only if $F_1 \ldots F_n \subseteq E$. The languages $F_1, \ldots, F_n$ are called the *terms* of the subfactorization. A term F_i is *maximal* if it can not be increased without violating the HL inclusion. A *factorization* of E is a subfactorization in which every term is maximal. A subfactorization $F'_1 \ldots F'_n$ of E *dominates* another subfactorization $F_1 \ldots F_n$ of E if $F_i \subseteq F'_i$ for all $1 \leq i \leq n$. A *factor* of E is any term of some factorization of E. A *left* (resp. *right*) *factor* of E is one which can be the leftmost (resp. rightmost) term in some factorization of E.

RHLs have finitely many factors. E.g., we can reason as follows to show that the right factors of an RHL E are finitely many: F is right factor of E iff there is a factorization $G\,F$ of E iff $F = \bigcap_{h \in G} \{h\}^{-1}E$ for some hedge language G. We noticed that $\{\{h\}^{-1}E \mid h \text{ a hedge}\}$ is a finite set of RHLs. Therefore, the right factors of E are intersections of RHLs taken from a finite set. Hence, they are RHLs (because RHLs are closed under intersection) and finitely many.

Note that F is a left factor of an HL E if and only if F^{s} is a right factor of the symmetric language E^{s}. This property enables to conclude that the set of right factors of an RHL is finite too, and to reduce the computation of left factors of an RHL to a computation of right factors of an RHL and vice versa.

From now on we consider only the problem of computing LSHs for the left factors of L when we know an LSH S made of $x_i = c_i + \sum_{k=1}^{m} \ell_{ik}\, x_k$ $(1 \leq i \leq m)$ with solution σ such that $\sigma(x_1) = L$. We tackle this problem in two steps: (1) Compute LSHs for all RHLs $L\{h\}^{-1}$ when h ranges over all hedges. (We saw already that this set of RHLs is finite. We call these RHLs the *right derivatives* of L); (2) Use the LSHs produced in step 1 to compute one LSH whose solution contains bindings to all possible intersections of right derivatives of L.

Let $\mathcal{X} = \{x_i \mid 1 \leq i \leq m\}$ and $\mathcal{I} := \{i_1, \ldots, i_s\} = \{i \mid x_1 \to^* x_i\}$ where $i_1 = 1$ and $\to^*$ is the reflexive-transitive closure of $\to$ defined by: $x_i \to x_j$ if x_j occurs at horizontal position in the right side of the i-th equation of S.

We have $x_j \doteq_\sigma c_j + \sum_{k=1}^{m} \ell_{jk}\, x_k$ for all $j \in \mathcal{I}$, and if we multiply all these relations to the right with $\{h\}^{-1}$, we obtain $y_j \doteq_\mu d_j(h) + \sum_{k=1}^{m} \ell_{jk}\, y_k$ where $d_j(h) = 1$ if $h \in [\![x_j]\!]_\sigma$ and $d_j(h) = 0$ otherwise, and μ extends σ with $\mu(y_j) := [\![x_j]\!]_\sigma \{h\}^{-1}$ for all $j \in \mathcal{I}$. Hence the LSHs for the right derivatives of L are

$$\begin{array}{ll} y_{i_j} = v_{i_j} + \sum_{k=1}^{m} \ell_{i_j k}\, y_k & (1 \leq j \leq s) \\ x_i = c_i + \sum_{k=1}^{m} \ell_{ik}\, x_k & (1 \leq i \leq m) \end{array} \tag{2}$$

with $(v_{i_1}, \ldots, v_{i_s}) \in \mathit{\Delta} := \{(d_{i_1}(h), \ldots, d_{i_s}(h)) \mid h \in \mathcal{H}(\mathit{\Sigma}, \emptyset)\}$. In order to compute the set $\mathit{\Delta}$, we define the relation $M \bowtie N$ for $M, N \in 2^{\mathcal{X}}$, with the reading "there exists a hedge h that belongs to $[\![x]\!]_\sigma$ for all $x \in M$ and does not belong to any $[\![x']\!]_\sigma$ when $x' \in N$." Then $(v_{i_1}, \ldots, v_{i_s}) \in \mathit{\Delta}$ if and only if there exist $M, N \in 2^{\mathcal{X}}$ such that $M \cup N = \{x_j \mid j \in \mathcal{I}\}$, $M \bowtie N$ holds, and $\{j \in \mathcal{I} \mid v_j = 1\} = \{j \in \mathcal{I} \mid x_j \in M\}$. Thus, in order to compute $\mathit{\Delta}$ it is sufficient to be be able to compute the pairs $\langle M, N\rangle \in 2^{\mathcal{X}} \times 2^{\mathcal{X}}$ for which $M \bowtie N$ holds.

It is easy to see that $M \bowtie N$ holds if and only if it can be inferred with

$$\frac{[\forall j \in J.\, c_j = 1 \wedge \forall k \in K.\, c_k = 0]}{\{x_j \mid j \in J\} \bowtie \{x_k \mid k \in K\}}$$

$$\frac{\{x_{1j} \mid j \in J\} \bowtie \{x_{1n} \mid n \in N_1\} \quad \{x_{2j} \mid j \in J\} \bowtie \{x_{2n} \mid n \in N_2\} \quad [\alpha]}{\{x_j \mid j \in J\} \bowtie \{x_k \mid k \in K\}}$$

where the side condition $[\alpha]$ of the second inference rule is

> there is an $a \in \mathit{\Sigma}$ such that $(\forall j \in J.\, a\langle x_{1j}\rangle\, x_{2j} \in \mathrm{rhs}(x_j))$ and
> $\{a\langle x_{1n}\rangle\, x_{2n} \mid n \in N\} = \{a\langle x\rangle\, x' \mid a\langle x\rangle\, x' \in \bigcup_{k \in \mathcal{K}} \mathrm{rhs}(x_k)\}$ and
> $\{N_1, N_2\}$ is a partition of N (that is, $N_1 \cup N_2 = N$ and $N_1 \cap N_2 = \emptyset$)

and the meaning of $\mathrm{rhs}(x_j)$ is as defined in Sect. 4. The first inference rule is valid because $\epsilon \in \bigcap_{j \in J} [\![x_j]\!]_\sigma \setminus \bigcup_{k \in K} [\![x_k]\!]_\sigma$, whereas the second inference rule is valid because of the existence of a hedge $a\langle h_1\rangle\, h_2 \in \bigcap_{j \in J} [\![x_j]\!]_\sigma \setminus \bigcup_{k \in K} [\![x_k]\!]_\sigma$. These inference rules are finitely branching and they constitute an inductive definition for the relation $M \bowtie N$ defined on a finite set of $2^m \times 2^m$ pairs. Therefore, these inference rules render a decision algorithm for the relation $M \bowtie N$, and this yields an algorithm for the computation of $\mathit{\Delta}$.

We have just seen how to compute LSHs for all right derivatives of L, and that these LSHs share the common structure of (2). Suppose we have computed p such LSHs $S_1, \ldots, S_p$ where every S_l is of the form

$$\begin{array}{ll} y^l_{i_j} = v^l_{i_j} + \sum_{k=1}^m \ell_{i_j k}\, y^l_k & \qquad (1 \leq j \leq s) \\ x_i = c_i + \sum_{k=1}^m \ell_{ik}\, x_k & \qquad (1 \leq i \leq m) \end{array}$$

with the specific set of variables $\mathcal{Y}_l = \{y^l_{i_1}, \ldots, y^l_{i_s}\}$ besides the set of variables $\mathcal{X}$ that is shared by all of them. Let's denote the unique solution of S_l by σ_l.

The left factors of L are the elements of the set $\{\bigcap_{l \in G} [\![y^l_1]\!]_{\sigma_l} \mid G \in 2^{\{1,\ldots,p\}}\}$. We consider the set of variables $\mathcal{Z} := \left\{\bigcap_{l \in G} y^l_{k_l} \,\middle|\, G \in 2^{\{1,\ldots,p\}} \wedge \forall l \in G.\, k_l \in \mathcal{I}\right\} \cup \{\bigcap_{i \in H} x_i \mid H \in 2^{\{1,\ldots,m\}}\}$ where variable names are identified modulo associativity, commutativity, and idempotency of intersection. We will construct an LSH LF with variables from $\mathcal{Z}$ whose unique solution μ satisfies the conditions: (c_1) $\mu(\bigcap_{l \in G} y^l_{k_l}) = \bigcap_{l \in G} [\![y^l_{k_l}]\!]_{\sigma_l}$ for every variable $\bigcap_{l \in G} y^l_{k_l}$ that occurs in LF; and (c_2) $\mu(\bigcap_{i \in H} x_i) = \bigcap_{i \in H} [\![x_i]\!]_\sigma$ for every variable $\bigcap_{i \in H} x_i$ that occurs in LF. Our main requirement is that variables of $\{\bigcap_{l \in G} y^l_1 \mid G \subseteq \{1, \ldots, p\}\}$ appear in LF. Then LF can be regarded as LSH for every left factor of L because every left factor of L is $[\![\bigcap_{l \in G} y^l_1]\!]_\mu$ for some $G \in 2^{\{1,\ldots,p\}}\}$, and a rearrangement of the equations of LF which places the equation for $\bigcap_{l\in G} y_1^l$ first is an LSH for the left factor $[\![\bigcap_{l\in G} y_1^l]\!]_\mu$. The equations of LF are constructed incrementally, by intersecting equations of the LSHs $S_1, \ldots, S_p$:

- For every $G \in 2^{\{1,\ldots,p\}}$ we intersect the first equations of the LSHs from the set $\{S_l \mid l \in G\}$. The intersection of any number of equations is the obvious generalization of the intersection operation of 2 equations described in Sect. 3. There are 2^p such intersections, and they will produce 2^p equations with variables $\bigcap_{l\in G} y_1^l$ in their left sides. We add these equations to LF.
- For every variable $\bigcap_{l\in G} y_{k_l}^l$ that occurs at horizontal position in some equation of LF, add (if missing) to LF the equation obtained by intersecting the equations of the set $\{k_l$-th equation of $S_l \mid l \in G\}$.
- For every variable $\bigcap_{i\in H} x_i$ that occurs in some equation of LF, add (if missing) to LF the equation produced by intersecting the equations of the set $\{i$-th equation of the LSH for $L \mid i \in H\}$.

This process terminates because $\mathcal{Z}$ is finite, so we can not add indefinitely equations to LF. We end up with LF being an LSH with properties (c_1) and (c_2).

6 Conclusion

LSHs are a representation of RHLs that is suitable for performing several operations that show up in the analysis and processing of XML. The algorithms described here indicate how the intersections, quotients, and the left and right factors of RHLs can be computed when using the LSH formalism. It should be mentioned, however, that there are also several operations for which LSHs are not a suitable representation, such as the computation of symmetric language.

References

1. V. M. Antimirov. Partial derivatives of regular expressions and finite automaton constructions. *Theoretical Computer Science*, 155:291–319, 1996.
2. J. H. Conway. *Regular Algebra and Finite Machines.* Mathematics series. Chapman and Hall, 1971.
3. H. Hosoya, J. Vouillon, and B. C. Pierce. Regular expression types for XML. *ACM Transactions on Programming Languages and Systems*, 27(1):46–90, 2005.
4. D. C. Kozen. *Automata and Computability.* Undergraduate Texts in Computer Science. Springer-Verlag New York, Inc., 1997.
5. M. Marin and T. Kutsia. Computational methods in an algebra of regular hedge expressions. RISC Report Series 09-03, RISC-Linz, March 2009.
6. M. Murata. Hedge automata: a formal model for XML schemata. http://www.xml.gr.jp/relax/hedge_nice.html, 1999.
7. M. Murata. Extended path expressions for XML. In *Proceedings of the 20th symposium on Principles of Database Systems (PODS'2001)*, pages 126–137, Santa Barbara, California, USA, 2001. ACM.
8. M. Murata, D. Lee, M. Mani, and K. Kawaguchi. Taxonomy of XML schema languages using formal language theory. *ACM Transactions on Internet Technology*, 5(4):660–704, 2005.