

Similarity relations and cover automata

Jean-Marc Champarnaud¹, Franck Guingne^{1,2} and Georges Hansel¹

¹ LIFAR, Université de Rouen,
{jmc, guingne, hansel}@dir.univ-rouen.fr

² XRCE, Xerox Research Center Europe, 38240 Meylan
franck.guingne@xrce.xerox.com

Abstract. Cover automata for finite languages have been much studied a few years ago. It turns out that a simple mathematical structure, namely similarity relations over a finite set of words, is underlying these studies. In the present work, we investigate in detail for themselves the properties of these relations beyond the scope of finite languages. New results with straightforward proofs are obtained in this generalized framework, and previous results concerning cover automata are obtained as immediate consequences.

1 Introduction

Let Σ be an alphabet and $\Sigma^{\leq l}$ be the subset of words of Σ^* whose length is not greater than the integer l . A relation over $\Sigma^{\leq l}$ is semi-transitive if, given three words $x, y, z \in \Sigma^{\leq l}$ such that $|x| \leq |y| \leq |z|$, transitivity holds when $x \sim y \wedge y \sim z$ or $y \sim x \wedge x \sim z$. In this paper, we present a general study of similarity relations over $\Sigma^{\leq l}$, i.e. relations that are reflexive, symmetrical and semi-transitive. We show in particular that right invariant similarity relations are recognized by semiautomata and we characterize minimal semiautomata recognizing a given relation.

We use these general properties to study cover automata for a finite language. Cover automata have been introduced by Câmpeanu, Sântean and Yu in [1]. A finite language L is said to be of order l if the length of a longest word in L is equal to l . A cover automaton for a language L of order l is a deterministic automaton \mathcal{A} such that $L(\mathcal{A}) \cap \Sigma^{\leq l} = L$. Checking word membership to L on a cover automaton for L only requires an additional test on the length of the word. Since covering generally reduces the size of an automaton [6], it is of practical interest to be able to compute a minimal cover automaton for L , that is a cover automaton with a minimal number of states. It is shown in [1] that a minimal cover automaton can be obtained from any cover automaton for L by merging states according to a state relation involving the right languages of the states. Minimality with respect to L comes from the properties of the similarity relation over $\Sigma^{\leq l}$ that is underlying the state relation. This word relation, called L -similarity, has been introduced by Kaneps and Freivalds [4] and Dwork and Stockmeyer [3].

In this paper, we show how a semiautomaton recognizing the L -similarity relation can be equipped with final states to yield a cover automaton for L . This leads to a characterization of minimal cover automata for a finite language.

Notice that several efficient algorithms have been designed for computing a minimal cover automaton, either from a deterministic automaton recognizing L , or from an arbitrary cover automaton for L . In [1], Câmpeanu, Sântean and Yu present an $O(n^4)$ time and space algorithm to minimize an n -state cover automaton for L . In [2], Câmpeanu, Păun and Yu provide an $O(n^2)$ time and space algorithm whose input is an n -state deterministic automaton recognizing L . In [5], Körner describes an Hopcroft-like algorithm with an $O(n \log n)$ time and $O(n)$ space complexity that works on both types of input.

Section 2 is devoted to a general study of similarity relations over $\Sigma^{\leq l}$ and Section 3 addresses right invariance property. The connexion between similarity relations and semiautomata is investigated in Section 4. The application of the study of similarity relations to the computation of a minimal cover automaton for a finite language is developed in Section 5.

2 Similarity relations over $\Sigma^{\leq l}$

Let l be an integer. In the following, $\Sigma^{\leq l}$ denotes the subset of Σ^* of words having a length not greater than l .

A relation \sim over $\Sigma^{\leq l}$ is *semi-transitive* iff for all x, y, z in $\Sigma^{\leq l}$ such that $|x| \leq |y| \leq |z|$, the following implications hold:

- (i) $x \sim y$ and $y \sim z \Rightarrow x \sim z$,
- (ii) $x \sim y$ and $x \sim z \Rightarrow y \sim z$.

A reflexive, symmetrical and semi-transitive relation is a *similarity relation*. In the following, the relation \sim is supposed to be a similarity relation over $\Sigma^{\leq l}$. Two words x and y are *similar* (resp. *dissimilar*) if $x \sim y$ (resp. $x \not\sim y$). A *similarity set* (resp. a *dissimilarity set*) is a subset of pairwise similar (resp. pairwise dissimilar) elements of $\Sigma^{\leq l}$. A dissimilarity set is *maximal* if its cardinality is maximal among dissimilarity sets. A partition of $\Sigma^{\leq l}$ whose all classes are similarity sets is called a *similarity partition*. A similarity partition is *minimal* if its cardinality is minimal among similarity partitions. Two similarity sets S and T are said to be *mergeable* if $S \cup T$ is a similarity set. Hence the partition resulting from merging two mergeable classes of a similarity partition is again a similarity partition.

An element $x \in \Sigma^{\leq l}$ is *minimal* if for all $y \in \Sigma^{\leq l}$, we have

$$y \sim x \Rightarrow |y| \geq |x|$$

We denote by M the set of all minimal elements of $\Sigma^{\leq l}$.

Proposition 1. 1) *The restriction of the relation \sim to M is an equivalence relation.*

2) *For all $x \in \Sigma^{\leq l}$, there exists at least one minimal element similar to x .*

Proof. 1) It follows from the very definition of minimal elements that two minimal similar elements have the same length. Consequently, by Condition (i), when restricted to M , the relation \sim is transitive.

2) Let $x \in \Sigma^{\leq l}$. Let y be an element of smallest length among all elements similar to x . It follows from Condition (i) that y is a minimal element.

Let us fix some notation. We denote by $\pi_M = \{M_1, \dots, M_k\}$ the partition of M in equivalence classes and by $C = \{c_1, \dots, c_k\}$ a cross-section of π_M , i.e. $c_i \in M_i$ for all $i = 1, \dots, k$. For all $x \in M$, let us denote by S_x the similarity set of all the elements similar to x . Finally, for all $i = 1, \dots, k$, let us set

$$T_i = S_{c_i} \setminus \bigcup_{j=1}^{i-1} S_{c_j} \text{ and } T'_i = S_{c_i} \setminus \bigcup_{j \neq i} S_{c_j}$$

Remark 1. It follows from Condition (i) that if x and x' are similar minimal elements, then $S_x = S_{x'}$. Moreover it follows from Proposition 1-2 that $\bigcup_{x \in M} S_x = \Sigma^{\leq l}$.

Proposition 2. 1) *The set C is a maximal dissimilarity set.*

2) *Any minimal similarity partition has k elements and $\{T_1, \dots, T_k\}$ is such a minimal similarity partition.*

Proof. 1) Being a cross-section of M , the set C is a dissimilarity set. Let D be any dissimilarity set. Suppose that $|D| > |C|$. Hence it follows from Proposition 1-2 that there exist two elements y and z in D similar to a same element c of C . Since c is a minimal element, $|y| \geq |c|$ and $|z| \geq |c|$ and therefore, by Condition (ii), we get that y and z are similar, a contradiction.

2) Let π be a similarity partition of $\Sigma^{\leq l}$. Different elements of C belong to different elements of π . Hence π has at least k elements. It remains only to observe that $\{T_1, \dots, T_k\}$ is a similarity partition (cf. Remark 1).

The following proposition gives a complete characterization of maximal dissimilarity sets.

Proposition 3. *Let D be a subset of $\Sigma^{\leq l}$. The following conditions are equivalent:*

1) *D is a maximal dissimilarity set.*

2) *$|D| = k$ and, for all $i = 1, \dots, k$, there exists one and only one element $d_i \in D$ such that $d_i \in T'_i$.*

Proof. 1) \Rightarrow 2) Since D is a maximal dissimilarity set, it follows from Proposition 2-1 that $|D| = |C| = k$. By Proposition 1-2, we can chose for all $d \in D$ a minimal element $f(d) \in C$ such that $f(d) \sim d$. Let d, d' be two elements of D and suppose that $f(d) = f(d')$. It follows from Condition (ii) that $d \sim d'$ and, since D is a dissimilarity set, we get that $d = d'$. Hence the mapping $d \rightarrow f(d)$ is one-to-one onto. Let $d_i = f^{-1}(c_i)$, $i = 1, \dots, k$. Then $d_i \sim c_i$ and $d_i \not\sim c_j$ for $j \neq i$ (otherwise we would get $d_i \sim d_j$). Hence $d_i \in T'_i$, $i = 1, \dots, k$, and 2) is

satisfied.

2) \Rightarrow 1) It suffices to observe that according to the definition of the sets T'_i , $i = 1, \dots, k$, we get that $D = \{d_1, \dots, d_k\}$ is a dissimilarity set.

Corollary 1. *Let D be a dissimilarity set. The following conditions are equivalent:*

- 1) D is a maximal dissimilarity set.
- 2) D is a cross-section of a similarity partition of $\Sigma^{\leq l}$.

Proof. 1) \Rightarrow 2) According to Proposition 3, $D = \{d_1, \dots, d_k\}$, with $d_i \in T'_i$ for all $i = 1, \dots, k$. Hence D is a cross-section of the similarity partition $\{T_1, \dots, T_k\}$.

2) \Rightarrow 1) Let $\pi = \{U_1, \dots, U_p\}$ be a similarity partition of $\Sigma^{\leq l}$ whose D is a cross-section. Since π is a similarity partition, we have $p \geq k$ and since D is a dissimilarity set, we have $p \leq k$. Hence $|D| = p = k$. We can assume that $c_i \in U_i$ for all $i = 1, \dots, k$ and denote by d_i the unique element of $D \cap U_i$. Since D is a dissimilarity set, we get that $d_i \sim c_j$ if and only if $i = j$. Hence $d_i \in T'_i$ for all $i = 1, \dots, k$ and D is a maximal dissimilarity set (cf. Proposition 3).

Lemma 1. *Let S and T be two similarity sets. Let s (resp. t) be one of the smallest elements of S (resp. T). The following conditions are equivalent:*

- 1) S and T are mergeable,
- 2) s and t are similar.

Proof. 1) \Rightarrow 2) is obvious. Let us prove that 2) \Rightarrow 1). Suppose that $|s| \leq |t|$. Let y be an element of S and z be an element of T . Since $|s| \leq |t| \leq |z|$, by Condition (i) we get that $s \sim z$. Consequently, since $|s| \leq |y|$ and $|s| \leq |z|$, by Condition (ii) we get that $y \sim z$. Hence S and T are mergeable.

Theorem 1. *Let π be a similarity partition of $\Sigma^{\leq l}$. The following conditions are equivalent:*

- 1) π is a minimal similarity partition.
- 2) π admits a maximal dissimilarity cross-section.
- 3) π admits a dissimilarity cross-section.
- 4) π cannot be reduced by merging elements,

Proof. 1) \Rightarrow 2) Since π is minimal, it has k elements (cf. Proposition 2) and consequently the set C is a maximal dissimilarity cross-section of π .

2) \Rightarrow 3) is obvious.

3) \Rightarrow 1) Let D be a dissimilarity cross-section of π . It follows from Corollary 1 that D is a maximal dissimilarity set. Hence D has k elements and π is minimal. Thus we have already shown that 1) \Leftrightarrow 2) \Leftrightarrow 3). The implication 1) \Rightarrow 4) is obvious and it follows from Lemma 1 that 4) \Rightarrow 3). The proof is complete.

3 Right invariant similarity relations

A similarity relation \sim over $\Sigma^{\leq l}$ is *right invariant* if $x \sim y \Rightarrow xz \sim yz$, for all $z \in \Sigma^*$ such that $|xz|, |yz| \leq l$. A similarity partition (U_i) is *right invariant* with respect to \sim if the conditions $x, y \in U_i$, $|xz| \leq l$, $|yz| \leq l$, and $xz \in U_j$ imply $yz \in U_j$.

Proposition 4. *Let \sim be a right invariant similarity relation over $\Sigma^{\leq l}$. Then there exists a minimal right invariant similarity partition.*

Proof. First we define a mapping $(c, a) \rightarrow c \cdot a$ from $C \times \Sigma$ to C by defining $c \cdot a$ as any element $c' \in C$ that is similar to the word ca . This mapping is then inductively extended to a mapping $(c, x) \rightarrow c \cdot x$ from $C \times \Sigma^*$ to C by setting

$$c \cdot x = \begin{cases} c & \text{if } x = \epsilon \\ (c \cdot y) \cdot a & \text{if } x = ya \end{cases}$$

Now we construct a partition of $\Sigma^{\leq l}$ denoted $\{U_1, \dots, U_k\}$, with $k = |C|$, by fixing, for all $x \in \Sigma^{\leq l}$, to which set U_i it belongs. Remark that the empty word ϵ belongs to C and we can suppose that $\epsilon = c_1$. Then we set

$$x \in U_i \Leftrightarrow \epsilon \cdot x = c_i$$

Let us first inductively check that $x \in U_i \Rightarrow x \sim c_i$ and hence that (U_i) is a similarity partition. By definition $\epsilon \in U_1$ and trivially $\epsilon = c_1 \sim c_1$. Suppose that $x = ya$ with $y \in U_j$ and $x \in U_i$. By the induction hypothesis, $y \sim c_j$. Since the relation \sim is right invariant we get that $x \sim c_j a$. On the other hand

$$\epsilon \cdot x = (\epsilon \cdot y) \cdot a = c_j \cdot a = c_i$$

By definition of $c_j \cdot a$, we have $c_j a \sim c_j \cdot a$. Thus we have $x \sim c_j a$ and $c_j a \sim c_j \cdot a$. But $|x| \geq |c_j a|$ and $|c_j a| \geq |c_j \cdot a|$. Hence $x \sim c_j \cdot a$, i.e. $x \sim c_i$.

Let us now check that the partition (U_i) is right invariant. Let $x \in U_i$ and let $a \in \Sigma$. Suppose that $c_i \cdot a = c_j$. Then

$$\epsilon \cdot xa = (\epsilon \cdot x) \cdot a = c_i \cdot a = c_j$$

Hence $U_i a \subset U_j$ and the partition (U_i) is right invariant.

Finally, since the partition (U_i) has $|C|$ elements, it is minimal.

4 Similarity relations and semiautomata

We assume that the reader is familiar with regular languages and automata theory [7].

Let $\mathcal{A} = (\Sigma, Q, q_-, \cdot)$ be a deterministic semiautomaton on the alphabet Σ . The *left language* of a state $q \in Q$ is defined by $\overleftarrow{L}(q) = \{x \in \Sigma^* \mid q_- \cdot x = q\}$. The set family $(\overleftarrow{L}(q))_{q \in Q}$ is a partition of $\Sigma^{\leq l}$.

A deterministic semiautomaton is a *similarity semiautomaton* for the relation \sim if for all $q \in Q$, $\overleftarrow{L}(q)$ is a similarity set. Such a semiautomaton is said to *recognize* the relation \sim . By definition, a similarity semiautomaton defines a similarity set partition of $\Sigma^{\leq l}$ and consequently it has at least $|\pi_M|$ states. Remark that the partition $(\overleftarrow{L}(q))$ cannot be an arbitrary similarity partition. Indeed, since for all $q \in Q$ and all $a \in \Sigma$, we have that $(\overleftarrow{L}(q) \cap \Sigma^{\leq l-1})a \subset \overleftarrow{L}(q \cdot a)$,

we get that the partition $(\overleftarrow{L}(q))$ is right invariant. We now will show that, conversely, to each right invariant similarity partition of $\Sigma^{\leq l}$, one can associate in a canonical way a similarity semiautomaton whose number of states is the index of the partition.

Let $(T_q)_{q \in Q}$ be a right invariant similarity partition. For all $x \in \Sigma^{\leq l}$, let us denote by q_x the element of Q such that $x \in T_{q_x}$ and for all $q \in Q$, let m_q be a word of minimal length in T_q . We define the mapping $(q, a) \rightarrow q \cdot a$ from $Q \times \Sigma \rightarrow Q$ by

$$q \cdot a = \begin{cases} q_{m_q a} & \text{if } m_q \in \Sigma^{\leq l-1} \\ q_\epsilon & \text{if } |m_q| = l \end{cases}$$

To complete the definition of the semiautomaton (Σ, Q, q_-, \cdot) , we set $q_- = q_\epsilon$.

Proposition 5. *Let $(T_q)_{q \in Q}$ be a right invariant similarity partition of $\Sigma^{\leq l}$. The semiautomaton (Σ, Q, q_-, \cdot) recognizes the relation \sim .*

Proof. a) Let us inductively prove that $q_- \cdot x = q_x$ for all $x \in \Sigma^{\leq l}$. If x is a letter $a \in \Sigma$, since $m_{q_-} = \epsilon$, one has

$$q_- \cdot a = q_{m_{q_-} a} = q_{\epsilon a} = q_a$$

Then we recursively get

$$q_- \cdot xa = (q_- \cdot x) \cdot a = q_x \cdot a = q_{m_{q_x} a}$$

Since x and m_{q_x} belong to the same element T_{q_x} of the partition (T_q) , using the invariance property of this partition, we get that $m_{q_x} a$ and xa belong to the same element $T_{q_{xa}}$. Hence we have

$$q_- \cdot xa = q_{m_{q_x} a} = q_{xa}$$

b) Let x and y be two words belonging to the same left language $\overleftarrow{L}(p)$, that is $q_- \cdot x = q_- \cdot y = p$. Then, using a), we get that $q_x = q_y = p$. Hence x and y belong to the same element T_p of the partition (T_q) and thus are similar. Consequently, the semiautomaton (Σ, Q, q_-, \cdot) recognizes the relation \sim .

Theorem 2. *Let \sim be a right invariant similarity relation over $\Sigma^{\leq l}$. Any similarity semiautomaton recognizing \sim has at least $|\pi_M|$ states and there exists a semiautomaton with $|\pi_M|$ states that recognizes \sim .*

Proof. By Proposition 4, there exists a minimal right invariant similarity partition $(T_q)_{q \in Q}$ with $|Q| = |\pi_M|$ and by Proposition 5, the associated semiautomaton (Σ, Q, q_-, \cdot) recognizes the relation \sim .

5 Cover automaton for a finite language

A finite language L is said to be *of order l* if l is the length of the longest word(s) in L . The notion of a cover automaton for a finite language L has been defined in [1] with the purpose of designing a compact representation of finite languages [6].

Definition 1. A cover automaton for a language L of order l is a deterministic finite automaton $\mathcal{C} = (\Sigma, Q, q_s, Q_+, \cdot)$ such that

$$L(\mathcal{C}) \cap \Sigma^{\leq l} = L$$

Definition 2. A cover automaton for a finite language L is minimal if it has a minimal number of states among the cover automata for L .

We now show that the underlying semiautomaton of a cover automaton for a language L recognizes the L -similarity relation introduced by Kaneps and Freivalds in [4] and Dwork and Stockmeyer [3].

Definition 3. Let L be a language of order l . Let x and y be two words of $\Sigma^{\leq l}$ and $h = \max\{|x|, |y|\}$. The relation \sim_L over $\Sigma^{\leq l}$, called L -similarity, is defined by:

$$x \sim_L y \Leftrightarrow (\forall t \in \Sigma^{\leq l-h}, xt \in L \Leftrightarrow yt \in L)$$

Lemma 2. The relation \sim_L is a right invariant similarity relation over $\Sigma^{\leq l}$.

Proof. The relation \sim_L is obviously reflexive and symmetrical. Let us show that it is semi-transitive. Let x, y, z be words of $\Sigma^{\leq l}$ such that $|x| \leq |y| \leq |z|$. We first check that $x \sim_L y$ and $y \sim_L z \Rightarrow x \sim_L z$. Let $t \in \Sigma^{\leq l}$ such that $|t| \leq l - |z|$. Since $y \sim_L z$, we have $yt \in L \Leftrightarrow zt \in L$. Since $|y| \leq |z|$ and $x \sim_L y$, it holds $xt \in L \Leftrightarrow yt \in L$. Consequently, $xt \in L \Leftrightarrow zt \in L$ and thus $x \sim_L z$. The proof of the second relation ($y \sim_L x$ and $x \sim_L z$) $\Rightarrow y \sim_L z$ is similar.

We now check that \sim_L is a right invariant relation. Let x and y be two words of $\Sigma^{\leq l}$ such that $\max\{|x|, |y|\} < l$. We have to prove that $x \sim_L y \Rightarrow xa \sim_L ya$ for all $a \in \Sigma$. Since $x \sim_L y$, we have $xt \in L \Leftrightarrow yt \in L$ for all $t \in \Sigma^{\leq l}$ such that $|t| \leq l - \max\{|x|, |y|\}$. Let us set $t = au$. It comes: $(xa)u \in L \Leftrightarrow (ya)u \in L$ for all $u \in \Sigma^{\leq l}$ such that $|u| \leq l - \max\{|x|, |y|\}$. Hence $xa \sim_L ya$.

Proposition 6. Let L be a language of order l .

- 1) Let \mathcal{C} be a cover automaton for L . Then the underlying semiautomaton of \mathcal{C} recognizes the relation \sim_L .
- 2) Conversely, a semiautomaton recognizing the relation \sim_L , when equipped with a convenient set of final states, is a cover automaton for L .

Proof. 1) Let $x, y \in \overleftarrow{L}(q)$, $h = \max\{|x|, |y|\}$. and let us show that $x \sim_L y$. Since $q_- \cdot x = q_- \cdot y$, we get that for all $z \in \Sigma^{\leq l-h}$, $q_- \cdot xz = q_- \cdot yz$. Hence $q_- \cdot xz \in Q_+ \Leftrightarrow q_- \cdot yz \in Q_+$. Consequently $xz \in L \Leftrightarrow yz \in L$ and $x \sim_L y$.

2) Conversely, let \mathcal{A} be a semiautomaton recognizing the relation \sim_L . Since $x \sim_L y \Rightarrow (x \in L \Leftrightarrow y \in L)$, we get that, for all $q \in Q$, either $\overleftarrow{L}(q) \subset L$ or $\overleftarrow{L}(q) \cap L = \emptyset$. Hence by setting $Q_+ = \{q \in Q \mid \overleftarrow{L}(q) \subset L\}$, the semiautomaton \mathcal{A} equipped with Q_+ is a cover automaton for L .

Theorem 3. Let L be a language of order l and \sim_L be the associated right invariant L -similarity relation over $\Sigma^{\leq l}$. Any cover automaton for the language L has at least $|\pi_M|$ states and there exists a cover automaton with $|\pi_M|$ states for L .

Proof. By Proposition 6-1 and Theorem 2, any cover automaton for L has at least $|\pi_M|$ states, and by Theorem 2 and Proposition 6-2, there exists a cover automaton with $|\pi_M|$ states.

References

1. C. Câmpeanu, N. Sântean, and S. Yu, *Minimal Cover-automata for Finite Languages*, Theoret. Comput. Sci. **267** (2001), 3–16.
2. C. Câmpeanu, A. Păun and S. Yu, *An Efficient Algorithm for Constructing Minimal Cover Automata for Finite Languages*, Intern. J. of Foundations of Comput. Sc., **13-1**(2002), 99–113.
3. C. Dwork and L. Stockmeyer, *A Time Complexity Gap for Two-Way Probabilistic Finite-State Automata*, SIAM J. on Computing, **19** (1990), 1011–1023.
4. J. Kaneps and R. Friedvalds, *Running Time to Recognize Non-Regular Languages by 2-Way Probabilistic Automata*, in ICALP'91, Lecture Notes in Computer Science, Springer-Verlag, 510(1991), 174–185.
5. H. Körner, *A Time and Space Efficient Algorithm for Minimizing Cover Automata for Finite Languages*, Int. J. of Foundations of Comput. Sci. **14** (2003), 1071–1086.
6. N. Sântean, *Towards a Minimal Representation for Finite Languages: Theory and Practice*, Master Thesis, The University of Western Ontario, 2000.
7. S. Yu, *Regular Languages*, in G. Rozenberg and A. Salomaa eds., Handbook of Formal Languages, Springer, Berlin, 1997.