Some investigations on similarity measures based on absent words

Giuseppa Castiglione, Sabrina Mantaci and Antonio Restivo
Dipartimento di Matematica e Informatica
Universitá di Palermo
Via Archirafi 34, 90123 Palermo, Italy
giuseppa.castiglione, sabrina.mantaci, antonio.restivo@unipa.it

Abstract. In this paper we investigate similarity measures based on minimal absent words, introduced by Chairungsee and Crochemore in [3]. They make use of a length-weighted index on a sample set corresponding to the symmetric difference $M(x)\triangle M(y)$ of the minimal absent words $M(x)$ and $M(y)$ of two sequences x and y, respectively. We first propose a variant of this measure by choosing as a sample set a proper subset $\mathcal{D}(x,y)$ of $M(x)\triangle M(y)$, which appears to be more appropriate for distinguishing x and y. From the algebraic point of view, we prove that $\mathcal{D}(x,y)$ is the base of the ideal generated by $M(x)\triangle M(y)$.

We then remark that such measures are able to recognize whether the sequences x and y share a common structure, but they are not able to detect the difference on the number of occurrences of such a structure in the two sequences. In order to take into account such a multiplicity, we introduce the notion of multifactor, and define a new measure that uses both absent words and multifactors. Surprisingly, we prove that this similarity measure coincides with a distance on sequences introduced by Ehrenfeucht and Haussler in [6], in the context of block-moves strategies. In this way, our result creates a non trivial bridge between similarity measures based on absent words and those based on the block-moves approach.

Keywords: Minimal absent words, similarity measures, sequence comparison.

1. Introduction

In recent years the concept of minimal absent word has been used to define similarity measures between two sequences. An advantage of this approach is that the set of minimal absent words of a sequence uniquely characterizes the sequence and, on the other hand, they are usually very short with respect to the sequence. Another motivation is that the number of minimal absent words of a sequence of length

Address for correspondence: {giuseppa.castiglione, sabrina.mantaci, antonio.restivo}@unipa.it.

n is linear in n, and then it is possible to compare sequences in time proportional to their lengths. In particular, in [3] Chairungsee and Crochemore introduced a measure of similarity between two sequences x and y making use of a *length-weighted index* on the symmetric difference $M(x)\triangle M(y)$ of the sets of minimal absent words $M(x)$ and $M(y)$ of x and y respectively.

Since, in some cases, the set $M(x)\triangle M(y)$ appears to be too large, in the same paper the authors propose to evaluate the length-weighted index on a *sample set*, i.e. the subset of $M(x)\triangle M(y)$ of words of length at most l, where l is a suitable integer. Further developments and an extension of the ideas of [3] can be found in [4]. An experimental study of different distance measures based on minimal absent words to analyze similarity/dissimilarity of sequences has been carried out in [10]. In the same paper the authors observe that "to the best of our knowledge there exists no attempt in the literature to identify the best index to employ on the minimal absent words set".

In this paper we introduce some new similarity measures between sequences based on minimal absent words. Our proposals come out from a theoretical investigation that analyzes some limitations and flaws of the measures proposed in [3].

The first similarity measure that we introduce is based on a choice of a sample set (subset of $M(x)\triangle M(y)$) different from the one considered in [3]. The criticism to the measure in [3] is that in the set $M(x)\triangle M(y)$ there are words absent both in x and in y (although they are minimal only for one of the two words). In our opinion, it is not appropriate to consider such words to distinguish x and y. So our first proposal is to evaluate the length-weighted index on the sample set

$$\mathcal{D}(x,y) = (F(x)\cap M(y)) \cup (F(y)\cap M(x)),$$

where $F(x)$ denotes the set of factors of x. The set $\mathcal{D}(x,y)$ contains words that are absent in one of the two words (x or y), but that are factors of the other one. In our proposal only the words of $\mathcal{D}(x,y)$ really contribute to separate x and y. Moreover, the set $\mathcal{D}(x,y)$ has some interesting algebraic properties: it is a subset of $M(x)\triangle M(y)$, and, moreover, it is the base of the ideal generated by $M(x)\triangle M(y)$. In other terms, the choice of $\mathcal{D}(x,y)$ as a sample set corresponds to eliminate those words of $M(x)\triangle M(y)$ that have a proper factor in the same set. Therefore, the deletion involves the longest words in $M(x)\triangle M(y)$. In this sense, our choice of the sample set is not very far from the one in [3], where the words up to a certain length are selected and, in addition, it appears more motivated and mathematically characterized.

The similarity measures proposed in [3], and also the one here proposed by taking $\mathcal{D}(x,y)$ as sample set, highlights whether two words share some similar structures. However, several examples, given in the paper, show that such measures are not able to detect whether such structures appear in the two words with a different multiplicity.

In particular, in the similarity measure that uses the sample set $\mathcal{D}(x,y)$, we take into account, for instance, how many elements of $M(y)$ are factors of the word x, but we do not consider how many times a single element of $M(y)$ occurs as factor of x. In order to take into account such a multiplicity, we introduce the notion of *multifactor*. Such a notion has an independent combinatorial interest and we introduce it in a more general context. A sequence $(v_1, v_2, \ldots, v_n)$ of elements (not necessarily distinct) in a language L is a L-multifactor of a word x if there exists a factorization

$$x = x_1v_1x_2v_2\cdots x_nv_nx_{n+1}$$

with $x_i \in \Sigma^*$, for $i = 1, 2, \ldots n+1$. The integer n is the *dimension* of the multifactor.

We introduce a new similarity measure by substituting the set $F(x)\cap M(y)$ with a $M(y)$-multifactor of x of maximal dimension (and, symmetrically, $F(y)\cap M(x)$ with a $M(x)$-multifactor of y). The similarity measure is related to such maximal dimension. Observe that the sequence corresponding to a $M(y)$-multifactor of x of maximal dimension is composed by elements of $F(x) \cap M(y)$, but each element could occur several times (or never) in the sequence.

We prove that this new similarity measure, based on minimal absent words and on their multiplicity, coincides with a distance on words introduced by Ehrenfeucht and Haussler in [6]. The idea of Ehrenfeucht and Haussler is very simple. The distance between two words corresponds to the minimal number of letters one need to delete in one word in such a way that the factors lying between two deleted characters are factors of the other word. The spirit of the definition of Ehrenfeucht and Haussler is very close to similarity measures based on block moves (cf. [8], [11], [12]). Indeed, the remaining factors of the first word can be thought as the results of block moves of the type described in [13].

In order to state the connection between our approach and that of Ehrenfeucht and Haussler, we prove that, given two words x and y, the maximal dimension of a $M(y)$-multifactor of x is equal to the number of letters on need to delete in x in such a way that the remaining factors are also factors of y.

In conclusion, the paper creates a non trivial link between the similarity measures based on absent words and those based on the block moves approach.

Let us, finally, remark that in this paper we are mainly focused on the combinatorial aspects of the similarity measures, and that the algorithmic point of view is not explicitly taken into account. However, the results of the paper show that all similarity measures considered in this paper are computable in linear time.

2. Definitions and notations

In this section we recall some fundamental notions and notations useful in the paper. Let Σ be a finite alphabet and Σ^* the set of the words over Σ. The set Σ^* is the free monoid generated by Σ with respect to the word concatenation and with the empty word ϵ the unit element. A set $I \subseteq \Sigma^*$ is said to be a (*two-sided*) *ideal* of Σ^* if for each $u \in I$ and $v \in \Sigma^*$ the two concatenations $uv, vu \in I$. This is equivalent to the condition $I = \Sigma^* I \Sigma^*$. The *base* of the ideal I is the minimal set B (with respect to the set inclusion) such that $I = \Sigma^* B \Sigma^*$.

Let v be a word of Σ^*, we say that u is a *factor* of v if there exist $z, w \in \Sigma^*$ such that $v = zuw$; if $z = \epsilon$ (resp. $w = \epsilon$) we say that u is a *prefix* (resp. a *suffix*) of v; if $u \neq v$ we say that u is a *proper factor* of v. If u is a prefix of v, i.e. $v = uz$, we denote $u^{-1}v = z$. In what follows we denote by $F(v)$ the set of factors of v. We say that a word u occurs in v if it is a factor of v. A subset P (resp. Q) of Σ^* is a *prefix set* (resp. *suffix set*) if none of its words is a prefix (resp. *suffix*) of another one. Any prefix (resp. suffix) set is a code, called *prefix code* (resp. *suffix code*). A prefix code is said to be *maximal prefix code* if it is maximal with respect to the inclusion of sets. For fundamentals in theory of codes cf. [2].

A word u is an *absent word* of v if it does not occur in v. An absent word is a *minimal absent word* if all its proper factors occur in v. We denote by $M(v)$ the set of minimal absent words of v.

A language $L \subseteq \Sigma^*$ is called *factorial* if it contains all the factors of its own words, whereas it is called *antifactorial* if no word in the language is a proper factor of another word in the language. In particular, for any word $v \in \Sigma^*$, $F(v)$ is a factorial language and $M(v)$ is antifactorial.

Remark that the complement of $F(v)$ (i.e. the set of the words that are not factors of v) is an ideal of Σ^* and $M(v)$ is its base. This allows to establish a duality between the sets $F(v)$ and $M(v)$ given by the relations (cf. [5]):

$$F(v) = \Sigma^* \setminus \Sigma^* M(v) \Sigma^*,$$

$$M(v) = \Sigma F(v) \cap F(v)\Sigma \cap (\Sigma^* \setminus F(v)).$$

This last relation comes from the remark that if $v \in \Sigma^*$, the word $u = a_1 a_2 \cdots a_n$, with $a_i \in \Sigma$ is a minimal absent word of v iff $u \notin F(v)$ and $a_1 \cdots a_{n-1}$, $a_2 \cdots a_n \in F(v)$.

3. Similarity measures based on sets of absent words

The notion of absent word plays an important role in many text processing algorithms. In particular, recently many papers dealing about string comparison use the notion of absent word in order to define distances between words. This definition is based on the intuition that two words, x and y are more distant if the set of the non common absent words is big and this set contains short words. In fact, long absent words of x that are present in y implicitly correspond to long common factors of x and y. Therefore long absent words must contribute less to the value of the distance of x and y.

This idea is formalized in a paper by Chairungsee and Crochemore [3] where the notion of length weighted index of a set is used in order to define a similarity/dissimilarity measure of two strings. The *length weighted index* is defined as the measure that associate to a set $X \subseteq \Sigma^*$ the quantity:

$$\mu(X) = \sum_{w \in X} \frac{1}{|w|^2}.$$

This measure is used in order to define a distance between two words x and y by taking the set $X = M(x) \triangle M(y)$, where $\triangle$ denotes the symmetric difference operator between two sets. The distance is

$$dist(x, y) = \mu(M(x) \triangle M(y)) = \sum_{w \in M(x) \triangle M(y)} \frac{1}{|w|^2}$$

Example 3.1. Let $x = cbaabdcb$ and $y = abcba$. Then,

$$M(x) = \{ac, ad, bb, bc, ca, cc, cd, da, db, dd, aaa, aba, bab, cbd, dcba\},$$

$$M(y) = \{aa, ac, bb, ca, cc, aba, bab, cbc, d\},$$

$$M(x) \triangle M(y) = \{d, aa, ad, bc, cd, da, db, dd, aaa, cbd, cbc, dcba\},$$

Therefore

$$(x, y) = 1 + \frac{7}{4} + \frac{3}{9} + \frac{1}{16} = \frac{453}{144}.$$

We remark that the distance between x and y is bigger if $M(x) \triangle M(y)$ contains short minimal absent words. In fact, longer minimal absent words contribute very few to the value of $dist$.

A different definition of distance is also given in the same paper (cf [3]). Let $M_l(x)$ (resp. $M_l(y)$) denote the set of minimal absent words of x (resp. y) having length at most l. Define:

$$dist_l(x,y) = \mu(M_l(x) \triangle M_l(y)) = \sum_{w \in M_l(x) \triangle M_l(y)} \frac{1}{|w|^2}$$

The choice of cutting away from the set $M(x) \triangle M(y)$ the words having length longer than a fixed bound comes from the observation that the lack of these words do not substantially affect the distance $dist(x,y)$, since there is an inverse relationship between $dist$ and the square of the lengths of the involved words. This choice, from the practical point of view, has an important advantage in terms of computation time.

Example 3.2. Let x and y be as in Example 3.1. If we consider $l = 2$, then $M_2(x) \triangle M_2(y) = \{d, aa, ad, bc, cd, da, db, dd\}$ and

$$dist_l(x,y) = 1 + \frac{7}{4} = \frac{11}{4}.$$

4. A new distance based on absent words

The distance $dist$ is computed as the measure μ of the set $M(x) \triangle M(y)$, whereas the distance $dist_l$ restricts the measure μ to the sample subset of the words of $M(x) \triangle M(y)$ having a bounded length. In this section we define a new distance also based on the measure μ, but applied on a different sample subset of $M(x) \triangle M(y)$. In particular the idea is to select in $M(x) \triangle M(y)$ only those words that really contribute to make the difference between x and y, i.e. those factors of x that are minimal absent words for y and viceversa. In other terms, we want the comparison of the two string x and y not to be influenced by those minimal absent words of y that do not occur in x. This idea is formally described as follows.

For all $x, y \in \Sigma^*$, we define the set:

$$D(x \leftarrow y) = F(x) \cap M(y)$$

i.e. the set of minimal absent words of y that are factors of x.

Given two words x and y we can define

$$\mathcal{D}(x,y) = D(x \leftarrow y) \cup D(y \leftarrow x) = (F(x) \cap M(y)) \cup (F(y) \cap M(x)).$$

In the following we prove some important properties of the set $\mathcal{D}(x,y)$.

Lemma 4.1. For all $x, y \in \Sigma^*$,

1. $D(x \leftarrow y) = \emptyset$ if and only if $x \in F(y)$;
2. $\mathcal{D}(x,y) = \emptyset$ if and only if $x = y$.
3. $\mathcal{D}(x,y) \subseteq M(x) \triangle M(y)$.
4. $\mathcal{D}(x,y)$ is antifactorial.

Proof:

1. Let $M(y) \cap F(x) = \emptyset$. Since $x \in F(x)$ then $x \notin M(y)$. Hence we have two possibilities: either $x \in F(y)$, and we are done; or x is absent in y but not minimal absent. In such a case x contains a minimal absent word of y as a factor. Hence, there exist $z, z_1, z_2 \in F(x)$ such that $x = z_1 z z_2$ and $z \in M(y)$. This means that $z \in M(y) \cap F(x)$. A contradiction.

   Conversely, suppose that x is a factor of y. We have that all the factors of x are factors of y too. It follows by definition that $M(y) \cap F(x) = \emptyset$.

2. We have that $\mathcal{D}(x, y) = (F(x) \cap M(y)) \cup (F(x) \cap M(y)) = \emptyset$ iff each set in the union is empty, that is iff x is a factor of y and y is a factor of x, i.e. $x = y$.

3. Let $z \in \mathcal{D}(x, y)$. Then either $z \in F(x) \cap M(y)$ or $z \in F(y) \cap M(x)$. Suppose $z \in F(x) \cap M(y)$ (the proof for the other case is symmetric). Then $z \in M(y)$ and $z \notin M(x)$, therefore $z \in M(y) \backslash M(x)$.

4. First, note that $\epsilon \notin \mathcal{D}(x, y)$ because it is never an absent factor i.e. $\epsilon \notin M(x)$ and $\epsilon \notin M(y)$. We have to prove that for all $z \in \mathcal{D}(x, y)$ none of its factor is in $\mathcal{D}(x, y)$. By contradiction suppose there exists a $z \in \mathcal{D}(x, y)$ and a factor z' of z such that $z' \in \mathcal{D}(x, y)$. Then either $z \in M(y) \cap F(x)$ or $z \in M(x) \cap F(y)$.

   Let us suppose that the first condition holds (in the other case we have an analogous proof). Then, in turn, either $z' \in M(y) \cap F(x)$ or $z' \in M(x) \cap F(y)$. In the first case we would have $z, z' \in M(y)$ and z' a factor of z, a contradiction since $M(y)$ is antifactorial. In the other case $z' \in M(x)$ and this is a contradiction since $z' \in F(z) \subseteq F(x)$.

□

Example 4.1. Let $x = cbaabdcb$ and $y = abcba$ as in Example 3.1. Then

$$M(x) \triangle M(y) = \{d, aa, ad, bc, cd, da, db, dd, aaa, cbd, cbc, dcba\},$$

$$\mathcal{D}(x, y) = \{d, aa, bc\}.$$

Note that as in the previous example the set $\mathcal{D}(x, y)$ is often much smaller than $M(x) \triangle M(y)$ and that $D(x, y)$ contains the words among the shortest of $M(x) \triangle M(y)$.

We remark that in Example 4.1, words like cd or aaa belonging to $M(x) \triangle M(y)$ should not contribute to measure the distance between x and y because they occur neither in x nor in y. So our idea is to define distance measures that do not depend on such words. In this scenario $\mathcal{D}(x, y)$ plays a fundamental role. Moreover, we prove that $\mathcal{D}(x, y)$ has an important algebraic property in relation to $M(x) \triangle M(y)$, as stated in the next theorem.

Theorem 4.1. Let $x, y \in \Sigma^*$. Then $\mathcal{D}(x, y)$ is the base of the ideal $\Sigma^*(M(x) \triangle M(y))\Sigma^*$.

Proof:
Since $\mathcal{D}(x, y)$ is antifactorial, in order to prove the statement it is sufficient to prove that any word $z \in (M(x) \triangle M(y)) \backslash \mathcal{D}(x, y)$ has a factor $z' \in \mathcal{D}(x, y)$. Since $z \in M(x) \triangle M(y)$, then

1. either $z \in M(x)$ and $z \notin M(y)$;
2. or $z \in M(y)$ and $z \notin M(x)$;

Let us consider case 1. The proof of case 2 is analogous.

By the hypothesis $z \notin D(x \leftarrow y) \cup D(y \leftarrow x)$ we have $z \notin M(x) \cap F(y)$. By conditions $z \in M(x)$, $z \notin M(y)$ and $z \notin M(x) \cap F(y)$ we deduce that $z \notin F(y)$. It means that z is absent in y but not minimal, therefore there exists a factorization $z = z_1 z' z_2$ with $z' \in M(y)$. Since $z \in M(x)$, its proper factors are in $F(x)$. Therefore $z' \in F(x) \cap M(y) = D(x \leftarrow y)$. □

In other terms, Theorem 4.1 states that $\mathcal{D}(x,y)$ can be obtained by deleting in $M(x) \triangle M(y)$ those words that have a proper factor in the same set. We are now ready to define a distance based on $\mathcal{D}(x,y)$:

$$\delta(x,y) = \mu(\mathcal{D}(x,y)) = \sum_{w \in \mathcal{D}(x,y)} \frac{1}{|w|^2} = \sum_{w \in D(x \leftarrow y)} \frac{1}{|w|^2} + \sum_{w \in D(y \leftarrow x)} \frac{1}{|w|^2}.$$

We remark that as in the case of $dist_l$, the distance δ takes into consideration elements among the shortest of $M(x) \triangle M(y)$ because they are elements of the base of the ideal $\Sigma^*(M(x) \triangle M(y))\Sigma^*$.

Example 4.2. Let $x = cbaabdcb$ and $y = abcba$. As shown in Example 4.1 $\mathcal{D}(x,y) = \{d, aa, bc\}$. Then:

$$\delta(x,y) = 1 + \frac{1}{2} = \frac{3}{2}$$

5. Similarity measure based on multiplicity of absent words

All the distances that we have seen so far are computed by applying a given measure (μ specifically, but one can use another definition of measure) on different sets of words associated to the sequences to compare. All these distances, highlight whether two sequences share a common ”structure”, but they are not able to detect whether such ”structures” appear in the two sequences with a different multiplicity.

In particular all of these distances are not sensible to the repetitions in strings.

The following example shows this limitation.

Example 5.1. For all $n \in \mathbb{N}$ let us consider the words $u_n = (ab)^n a$. It is easy to verify that for any $n \in \mathbb{N}$, $M(u_n) = \{aa, bb, (ba)^n b\}$ and then

$$M(u_1) \triangle M(u_n) = \{bab, (ba)^n b\}.$$

Therefore

$$dist(u_1, u_n) = \frac{1}{9} + \frac{1}{(2n+1)^2}.$$

We remark that, the value of $dist$ has an inverse relation with the number of repetitions. This means that, for instance, $dist(u_1, u_{100}) < dist(u_1, u_2)$, that is, according to $dist$, the word $(ab)^2 a$ is farther from aba than $(ab)^{100} a$ where the factor ab is repeated 100 times, whereas common sense would suggest the converse.

On the other side we can see that

$$D(u_1 \leftarrow u_n) = \emptyset \text{ and } D(u_n \leftarrow u_1) = \{bab\}$$

(no minimal absent words of u_n occur in u_1 and bab is the unique minimal absent word of u_1 that occurs in u_n). Therefore for any $n \in \mathbb{N}$, $\mathcal{D}(u_1, u_n) = \{bab\}$ and

$$\delta(u_1, u_n) = \frac{1}{9}$$

i.e. $\delta(u_1, u_n)$ depends neither from the relative lengths of the two words nor from the number of repetitions of the factor ab. For instance both the words $(ab)^2a$ and $(ab)^{100}a$ (where ab is repeated 100 times) have the same distance from aba. This shows that δ is able to detect that for all $n \in \mathbb{N}$, u_1 and u_n share a common "structure" (a and b alternate) but is not able to detect that such an alternation occurs with different multiplicities in the two sequences.

In this section we introduce a method of comparison between strings that captures information about the number of repetitions of a given factor. In particular, in the computation of the distance δ we do not take into account the multiplicity of the presence of any $v \in M(y)$ as a factor of x, i.e. the number of times that v occurs in x. In order to refine the similarity measure we can consider also these multiplicities. In what follows we introduce a method of comparison of two strings x and y that includes multiplicities of absent words of y as factors of x. We begin by giving the definition of multifactor of a word.

Definition 5.1. Let $\underline{v} = (v_1, v_2, \ldots, v_n)$ be a sequence of (not necessarily distinct) elements of Σ^*. We say that $\underline{v}$ is a *multifactor* of a word $x \in \Sigma^*$ if there exist $x_1, x_2, \ldots, x_{n+1} \in \Sigma^*$ such that $x = x_1v_1x_2v_2 \cdots x_nv_nx_{n+1}$.

The sequence of words $x_1, x_2, \ldots, x_n, x_{n+1}$ specify a *(non overlapping) occurrence of the multifactor $\underline{v}$ in x*, and is called a *context*. We say that n is the *dimension* of the multifactor. Multifactors of dimension 1 are the factors of x.

Example 5.2. Let $x = abbabbaaabb$. Then $\underline{v} = (bab, aaa, b)$ is a multifactor of x that has two different occurrences defined by the following contexts ab, b, ϵ, b, and ab, b, b, ϵ.

Remark 5.1. The notion of multifactor, althoug similar, is different from the one of (*scattered*) *subword* (cf. [9, Chap. 6]). In fact a scattered subword of a word x is a (single) word obtained from x by deleting some of its letters, whereas a multifactor is a sequence of words $\underline{v} = (v_1, v_2, \ldots, v_n)$ that can occur in the specified order from left to right as non overlapping factors of x. The concatenation of words in $\underline{v}$ give rise to a scattered subword of x, but the same scattered subword can be obtained from different multifactors. For instance if $x = abbabbaaabb$, then (bab, aab) and (ba, baa, b) are two different multifactors of x corresponding to the same scattered subword $babaab$ of x.

In the special case where $\underline{v} = (v_1, v_2, \ldots, v_n)$, with $v_1 = v_2 = \cdots = v_n = v$, the search of the multifactor $\underline{v}$ in a word w corresponds to the search of the distinct non overlapping occurrences of the factor v in w (cf. [1], [7]).

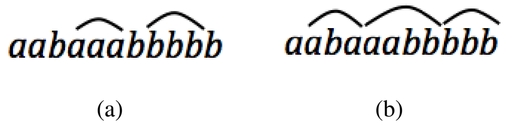


Figure 1. (a) A $M(y)$-multifactor, (b) a maximal $M(y)$-multifactor; with $x = aabaaabbbbb$ and $y = abbaab$.

If $L \subseteq \Sigma^*$ is a language, denote by $(L)^n$ the set of sequences $(v_1, v_2, \ldots, v_n)$ with $v_i \in L$ for $i = 1, 2, \ldots n$. Remark that $(L)^n$ denotes a set of sequences of words in L whereas L^n denotes a set of words, obtained by concatenating n elements of L.

If $\underline{v} \in (L)^n$ for some $n \in \mathbb{N}$ is a multifactor of a word $x \in \Sigma^*$, then $\underline{v}$ is called a *L-multifactor of* x.

We say that $\underline{v}$ is a *maximal L-multifactor* if it has a maximal dimension, i.e. any other $\underline{u} \in (L)^m$ multifactor of x is such that $m \leq n$.

Given $x, y \in \Sigma^*$, we are interested in $M(y)$-multifactors of x, i.e. multifactors of x composed by minimal absent words of y.

Definition 5.2. For $x, y \in \Sigma^*$ we denote by $amf(x \leftarrow y)$ the dimension of a maximal $M(y)$-multifactor of x.

Example 5.3. Let $x = aabaaabbbbb$ and $y = abbaab$. Then $M(y) = \{aaa, aabb, aba, bab, bbb\}$ as one can verify. As notations, if the factorization $x = x_1 v_1 x_2 v_2 .....x_n v_n x_{n+1}$ determines an occurrence of the $M(y)$-multifactor $(v_1, v_2, ..., v_n)$, we place an arc over every v_i in the above factorization. Figure 1(a) shows that (aaa, bbb) is a $M(y)$-multifactor of x and Figure 1(b) shows that $(aba, aabb, bbb)$ is a maximal $M(y)$-multifactor.

As Example 5.3 shows, a $M(y)$-maximal multifactor does not necessarily contain all the elements of $D(x \leftarrow y)$. Furthermore, a multifactor can have many occurrences. For example, let us consider $x = a^{12}$ and $y = a^4$. One has that $M(y) = \{a^5\}$ and the $M(y)$-multifactor (a^5, a^5) have many contexts such as, for instance, aa, ϵ, ϵ and a, a, ϵ.

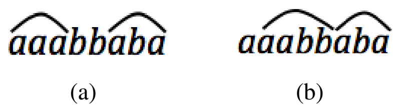


Figure 2. Two maximal $M(y)$-multifactors of $x = aaabbaba$ where $y = abbaab$.

Note that a word x can have many different maximal $M(y)$-multifactors, as shown in Figure 2.

The following theorem holds.

Theorem 5.1. Let $x, y \in \Sigma^*$ and let $\underline{v} = (v_1, v_2, \ldots, v_n)$ be a maximal $M(y)$-multifactor of x. Then, for any factorization $x = x_1 v_1 \cdots x_n v_n x_{n+1}$ one has that $x_i \in F(y)$, for $i = 1, 2, \ldots n + 1$.

Proof:
Let $\underline{v} = (v_1, v_2, \ldots, v_n)$ be a maximal $M(y)$-multifactor of x and $x = x_1v_1 \cdots v_{i-1}x_iv_i \cdots x_nv_nx_{n+1}$. If for some i, $x_i \notin F(y)$, then it is absent in y. Therefore, there exists $z \in M(y)$ and two words $u, w \in \Sigma^*$ such that $x_i = uzw$. Then $x = x_1v_1 \cdots v_{i-1}u\, z\, w\, v_i \cdots x_nv_nx_{n+1}$, i.e $(v_1, \ldots v_{i-1}, z, v_i, \ldots v_n)$ is a $M(y)$-multifactor of x with a greater dimension than the maximal one, a contradiction. □

The converse is not true, as shown in Figure 1(a), where an occurrence of the non maximal $M(y)$-multifactor (aaa, bbb) is shown. The words aab and b of its context are in $F(y)$.

Given $M(y)$ as input, we can find a particular $M(y)$-multifactor $\underline{v}_g = (v_1, v_2, \ldots v_n)$ of x by using a greedy strategy. The following greedy algorithm outputs the $M(y)$-multifactor $\underline{v}_g$ and the context $x_1, x_2, \ldots x_n$ that specify the occurrence of $\underline{v}_g$.

```
                    GREEDY-MF($x \leftarrow y$)

1. Scan $x$ from left to right until the leftmost factor $v_1$ in $M(y)$ is found or the end of $x$ is reached.
    (a) Determine the prefix $x_1$ of $F(y)$ (eventually empty) that occurs before $v_1$;
    (b) RETURN $(x_1, v_1)$;
2. GREEDY-MF($(x_1v_1)^{-1}x \leftarrow y$).
```

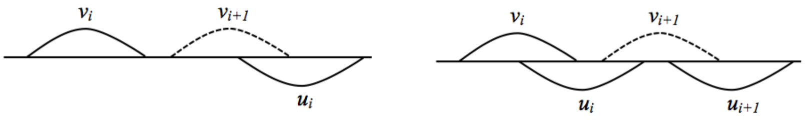
Theorem 5.2. The $M(y)$-multifactor $\underline{v}_g$ of x obtained by **GREEDY-MF**$(x \leftarrow y)$ is maximal.

The proof of Theorem 5.2 is a consequence of the following

Lemma 5.1. Let $x, y \in \Sigma^*$, $\underline{v}_g = (v_1, \ldots v_n)$ and $x = x_1v_1 \cdots x_nv_nx_{n+1}$ be the greedy factorization of x. Let $\underline{u} = (u_1, \ldots, u_m)$ be another $M(y)$-multifactor of x with $x = z_1u_1 \cdots z_mu_mz_{m+1}$ another factorization. Then for all $i \geq 1$, $x_1v_1 \cdots x_iv_i$ is a prefix of $z_1u_1 \cdots z_iu_i$.

Proof:
The proof is by induction on i.

- if $i = 1$ then x_1v_1 is a prefix of z_1u_1, since for the greedy strategy v_1 is the leftmost factor of x in $M(y)$.
- Assume by induction that $x_1v_1 \cdots x_iv_i$ is a prefix of $z_1u_1 \cdots z_iu_i$. We first remark that no factor in $\underline{v}_g$ can be a proper factor of $\underline{u}$ and viceversa, otherwise $M(y)$ would not be antifactorial. We have two possibilities:
    - u_i starts after the end of v_i (see Figure 3(a)). In such a case, because of the left-to right greedy strategy, we have that v_{i+1} begins before u_i or, at most, it coincides with it. It follows that v_{i+1} ends before at the end of u_i (or a most coincides with it). Then $x_1v_1 \cdots v_{i+1}$ is a prefix of $z_1u_1 \cdots z_iu_i$, and the claim follows.
    - v_i ends after the beginning of u_i and before the end of u_i (see Figure 3(b)). In such a case v_{i+1} ends before or the end of u_{i+1}, or at most it coincides with it.



(a) (b)

Figure 3. The two cases of proof of Lemma 5.1.

□

We recall that the set of minimal absent words of a word can be computed in linear time [5]. Since also **GREEDY-MF** runs in linear time we have:

Corollary 5.1. The value $amf(x \leftarrow y)$ can be computed in linear time.

In order to generalize the distance measures $dist$ and δ introduced in the previous section, this time we want to apply the measure μ to multifactors. This will stress the remark that an element of $M(y)$ ($M(x)$, resp.) that appears several times as factor of x (y, resp.) needs to have a higher weight in the computation of the distance than an element that appears only once.

Let $\underline{v}_g = (v_1, v_2, \dots v_k)$, with $k = amf(x \leftarrow y)$, be the maximal $M(y)$-multifactor of x and $\underline{u}_g = (u_1, u_2, \dots u_h)$, with $h = amf(y \leftarrow x)$, the maximal $M(x)$-multifactor of y, output by **GREEDY-MF**$(x \leftarrow y)$ and **GREEDY-MF**$(y \leftarrow x)$, respectively. Then we define:

$$\gamma(x, y) = \sum_{i=1}^{amf(x \leftarrow y)} \frac{1}{|v_i|^2} + \sum_{i=1}^{amf(y \leftarrow x)} \frac{1}{|u_j|^2}$$

As one can remark, the definition of the distance measure γ takes into account both the lengths and the multiplicities of the factors in v_g and in u_g.

Example 5.4. Consider the words $u_n = (ab)^n a$ as in Example 5.1. **GREEDY-MF**$(u_n \leftarrow u_1)$ returns the multivector $(bab, \dots, bab)$ of dimension $\lfloor \frac{n}{2} \rfloor$. Then

$$\gamma(u_1, u_n) = \sum_{i=1}^{\lfloor n/2 \rfloor} \frac{1}{|bab|^2} = \frac{1}{9} \lfloor \frac{n}{2} \rfloor.$$

Remark that, differently from distances $dist$ and δ, the distance γ strongly depends on the number of iterations of ab.

Remark 5.2. We remark that, in order to have a well defined distance, we need to consider the greedy maximal multifactors and not *any* maximal multifactors. In fact this measure takes into account not only the dimension of the maximal multifactor, but also the lengths of the single factors in it; then a different maximal multifactor could involve factors with different lengths, giving rise to a different value of γ.

The choice of using the multifactor output by **GREEDY-MF** is mainly motivated by the fact that the algorithm runs in linear time. However, we expect that different choices for maximal multifactors do not affect much the value of γ; analytical comparisons can be done in a future work. Another motivation is that the greedy strategy allows to find a connection between the similarity measure $amf(x \leftarrow y)$ and a similarity measure considered in the next section.

Remark 5.3. Although γ is influenced by the length of the words in the maximal multifactor and amf considers only the number of such words, both the measures give coherent information about similarity/dissimilarity. Indeed, as the dimension of the maximal multifactor v_g grows, the average length of the words v_i $(i = 1, \dots, amf)$ get shorter, hence the value of γ grows too.

However, by Remark 5.2, γ depends on the particular choice of the maximal multifactor, whereas amf is the same for all maximal multifactors. For such a reason it appears to be preferable to use amf for defining a similarity measure.

6. Ehrenfeucht-Haussler distance

In a paper issued in 1988 [6] Ehrenfeucht and Haussler introduced a method to compare two strings x and y, trying to capture non local similarities of two words. This method is quite different from the ones that we have seen so far, but it turns out that it also implicitly depends on minimal absent factors of x that appear as factor in y and viceversa. In this section we connect these comparison methods, apparently so different.

Informally speaking, the comparison method introduced by Ehrenfeucht and Haussler consists in counting the minimal number of letters one need to delete in x in such a way that the factors lying between two deleted characters are factors of y. We first recall some definitions introduced in the paper.

Given a nonempty word $x = a_1 a_2 \cdots a_n$ where $a_i \in \Sigma$ for $i = 1, \dots, n$, we say that x has n places, denoted by the numbers from 1 to n. A *marking* S of x is a (possibly empty) subset of k places of x. The set $\{a_i : i \in S\}$ is called the *set of S-marked letters* of x. For any $y \in \Sigma^*$, we say that S *makes x compatible with y* if any factor of x that does not contain any S-marked letters is also a factor of y.

Definition 6.1. For $x, y \in \Sigma^*$, $diff(x \leftarrow y)$ is the minimal number of places in any marking of x that makes x compatible with y.

In the following example we denote a marked symbol by underlining it. The minimal marking is not unique, as shown in the following:

Example 6.1. Let $x = aabaaabbbbb$ and $y = abbaab$ then a marking that makes x compatible with y is $aab\underline{a}aab\underline{b}bb\underline{b}$. Other different possible markings are $a\underline{a}baa\underline{a}bb\underline{b}bb$ or $aa\underline{b}aa\underline{a}bb\underline{b}bb$.

Definition 6.2. For $x, y \in \Sigma^*$, we denote by $S_l(x \leftarrow y)$ (or simply S_l, if x and y are understood by the context) the unique marking of x such that $x = x_1 m_1 x_2 m_2 \cdots x_k m_k x_{k+1}$, where, for $1 \le i \le k$, m_i is a S_l-marked letter of x, for $1 \le i \le k+1$ x_i is a factor of y, and $x_i m_i$ is not a factor of y. $S_r(x, y)$ (or simply S_r) is defined in the same way, except that we demand that $m_i x_{i+1}$ is not a factor of y, $1 \le i \le k$. In either case, when x is a factor of y, the definition holds vacuously with $k = 0$. In this case $S_l = S_r = \emptyset$.

The marking $S_l(x \leftarrow y)$ of x can be found in linear time by scanning x from left to right until a character a is found such that the prefix pa is not in $F(y)$ whereas p is in $F(y)$. Hence we mark a and repeat the process for the remaining suffix until the end of the string.

By scanning x from right to left, we can symmetrically define the procedure and obtain $S_r(x \leftarrow y)$.

In [6], Ehrenfeucht and Haussler prove the following

Theorem 6.1. Let $x, y \in \Sigma^*$. Then $|S_l(x \leftarrow y)| = |S_r(x \leftarrow y)| = diff(x \leftarrow y)$.

Hence, the greedy strategy described above produces the optimal solution to the problem of minimal marking.

Moreover in the same paper Ehrenfeucht and Haussler use $diff(x \leftarrow y)$ and $diff(y \leftarrow x)$ in order to define a metric on the set of strings over a given alphabet, as follows

$$d_{EH}(x, y) = \log(diff(x \leftarrow y) + 1)(diff(y \leftarrow x) + 1))$$

In the last part of this section we would like to show a connection between the size of the maximal $M(y)$-subfactor of x and the size of the minimal marking of x compatible with y. The following theorem establishes a bridge between the two concepts:

Theorem 6.2. Let $x, y \in \Sigma^*$, then

$$diff(x \leftarrow y) = amf(x \leftarrow y)$$

In order to prove Theorem 6.2, we introduce the following notion:

Definition 6.3. A word u is a *right minimal absent word* of y if it is absent in y and all its proper prefixes occur in y.

An equivalent definition is that the word $u = a_1 a_2 \cdots a_n$ is a right minimal absent word in $y \in \Sigma^*$, iff $u \notin F(y)$ and $a_1 a_2 \cdots a_{n-1} \in F(y)$.

In a symmetric way we can define left minimal absent words.

We denote by $RM(y)$ (resp. $LM(y)$) the set of all right (resp. left) minimal absent words of y. Trivially, $M(y) = RM(y) \cap LM(y)$.

Proposition 6.1. Let $y \in \Sigma^*$. Then:

$$RM(y) = \Sigma^* M(y) \backslash \Sigma^* M(y) \Sigma^+$$

Proof:
We prove first the inclusion $RM(y) \subseteq \Sigma^* M(y) \backslash \Sigma^* M(y) \Sigma^+$.

Let $u \in RM(y)$. We prove by induction on $|u|$ that $u \in \Sigma^* M(y)$.

- If $|u| = 1$, then $u \in M(y)$.
- Let $|u| > 1$. If $u \in M(y)$ we are done. In the other case, suppose that for any $u' \in RM(y)$ with $|u'| < |u|$, the statement is true. Since $|u| > 1$ and $u \in RM(y)$, then $u = bu'$ with $b \in \Sigma$ and $u' \in \Sigma^*$. Then $u' \notin F(y)$ (otherwise u would be in $M(y)$), then $u' \in RM(y)$ and $|u'| < |u|$. Therefore by inductive hypothesis $u' = v'w'$ with $v' \in \Sigma^*$ and $w' \in M(y)$. Then $u = bv'w'$ with $bv' \in \Sigma^*$ and $w' \in M(y)$.

Now we prove that $u \notin \Sigma^* M(y)\Sigma^*$. By contradiction suppose that there exist $z \in M(y)$, $z_1 \in \Sigma^*$, $z_2 \in \Sigma^+$ such that $u = z_1 z z_2$. But by definition of $RM(y)$ every proper prefix of u (and therefore all of its own factors, such as z) must be in $F(y)$. A contradiction.

Now we prove that $\Sigma^* M(y) \backslash \Sigma^* M(y)\Sigma^+ \subseteq RM(y)$.

Let $u = a_1 a_2 \cdots a_n \in \Sigma^* M(y) \backslash \Sigma^* M(y)\Sigma^+$, then $u \notin F(y)$ and $a_1 \cdots a_{n-1} \in F(y)$ because otherwise it is absent i.e. it contains an element of $M(y)$. This is a contradiction since $u \notin \Sigma^* M(y)\Sigma^+$. Then $u \in RM(y)$. □

The above proposition states, in other terms, that for all $y \in \Sigma^*$, $RM(y)$ is a semaphore code (cf. [2]). It is known that *semaphore codes* are maximal prefix codes, then any word $x \in \Sigma^*$ can be univocally factorized as:

$$x = y_1 y_2 \cdots y_m z$$

where $y_1, y_2, \ldots, y_m \in RM(y)$ and z is a prefix of a word in $RM(y)$. We call this factorization the *prefix factorization* of x in $RM(y)$.

The proof of the following lemma is straightforward

Lemma 6.1. The left-to-right marking that makes x compatible with y coincides with the marking of the last letter of each word in the prefix factorization of x in $RM(y)$.

Furthermore, since each $y_i \in RM(y)$, by Proposition 6.1, it can be written as $x_i u_i$ with $x_i \in \Sigma^*$ and $u_i \in M(y)$, then the prefix factorization induces a factorization of x as:

$$x = x_1 u_1 x_2 u_2 \cdots x_k u_k x_{k+1},$$

where $x_{k+1} = z$. The sequence $(u_1, u_2, \ldots u_n)$ is the $M(y)$-multifactor of x induced by the prefix factorization in $RM(y)$. The following lemma is a consequence of the above remarks and Proposition 6.1.

Lemma 6.2. The $M(y)$-multifactor of x output by **GREEDY-MF**$(x \leftarrow y)$ coincides with the $M(y)$-multifactor induced by the prefix factorization in $RM(y)$.

Proof of Theorem 6.2:
On one hand by Lemma 6.1 the prefix factorization of x in $RM(y)$ induces the Ehrenfeucht-Haussler marking. That is $dist(x \leftarrow y)$ corresponds to the number of factors in $RM(y)$ in such a factorization. On the other hand by Lemma 6.2 the number of such factors coincides with the number of element in the maximal multifactor v_g, which is equal to $amf(x \leftarrow y)$.

References

[1] Alberto Apostolico and Franco P. Preparata. Data structures and algorithms for the string statistics problem. *Algorithmica*, 15(5):481–494, 1996.

[2] Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata (Encyclopedia of Mathematics and Its Applications).* Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[3] Supaporn Chairungsee and Maxime Crochemore. Using minimal absent words to build phylogeny. *Theor. Comput. Sci.*, 450:109–116, 2012.

[4] Panagiotis Charalampopoulos, Maxime Crochemore, Gabriele Fici, Robert Mercas, and Solon P. Pissis. Alignment-free sequence comparison using absent words. *Inf. Comput.*, 262(Part):57–68, 2018.

[5] Maxime Crochemore, Filippo Mignosi, and Antonio Restivo. Automata and forbidden words. *Inf. Process. Lett.*, 67(3):111–117, 1998.

[6] Andrzej Ehrenfeucht and David Haussler. A new distance metric on strings computable in linear time. *Discrete Applied Mathematics*, 20(3):191–203, 1988.

[7] Sahar Hooshmand, Paniz Abedin, M. Oguzhan Külekci, and Sharma V. Thankachan. Non-overlapping indexing - cache obliviously. In *Annual Symposium on Combinatorial Pattern Matching, CPM 2018, July 2-4, 2018 - Qingdao, China*, pages 8:1–8:9, 2018.

[8] Haim Kaplan and Nira Shafrir. The greedy algorithm for edit distance with moves. *Information Processing Letters*, 97(1):23 – 27, 2006.

[9] M. Lothaire. *Combinatorics on Words*. Addison-Wesley, 1983.

[10] Mohammad Saifur Rahman, Ali Alatabbi, Maxime Crochemore, and Mohammad Sohel Rahman. Absent words and the (dis)similarity analysis of dna sequences: An experimental study. *BMC Research Notes.*, 9:186 450:1–8, 2016.

[11] Dana Shapira and James A. Storer. Edit distance with move operations. *Combinatorial Pattern Matching. CPM 2002. Lecture Notes in Computer Science,*, 2373, 2002.

[12] Dana Shapira and James A. Storer. Edit distance with move operations. *Journal of Discrete Algorithms*, 5(2):380 – 392, 2007. 2004 Symposium on String Processing and Information Retrieval.

[13] Walter F. Tichy. The string-to-string correction problem with block moves. *ACM Trans. Comput. Syst.*, 2(4):309–321, 1984.