Fuzzy words

Anton Černý
Department of Information Science, Kuwait University
anton.cerny@ku.edu.kw

Abstract

We propose to study words over a partially ordered alphabet with different weaker kinds of matching of two words, based on the partial order relation. Fuzzy words are a generalization (refining) of the concept of partial words, extensively studied within problems arising, e.g., in DNA sequencing. We provide a few basic results on position of fuzzy closures of languages within the Chomsky hierarchy, periodicity and conjugacy of fuzzy words, as well as adaptation to fuzzy words of the extended Parikh matrix morphism - a subword counting tool.

Keywords: fuzzy word, partial word, period, conjugacy, Chomsky hierarchy

1 Introduction

In computing sciences, to be able to process any kind of information, the information has to be encoded by a sequence of symbols. A finite sequence of symbols from some (finite) alphabet is called a word (string). However, sometimes (e.g., in DNA sequencing) some piece of information may be missing or hidden. This can be manifested by positions denoting missing symbols in a word. Thus, instead of complete words, only partial words are to be considered (see, e.g.([**?**]) for more details). The missing symbols are usually denoted as " holes" and denoted by a special symbol, say $\diamond$ (not belonging to the alphabet Σ under consideration). The position in the word denoted by $\diamond$ may contain any symbol from the alphabet Σ. In this sense, the symbol $\diamond$ stands for the whole alphabet Σ. What if we know that the missing symbol in a word is just from a limited subset of Σ? We need another special symbol for this subset, different from $\diamond$. It is therefore resonable to consider alphabets with several additional special symbols, corresponding to subsets of Σ. In fact, a symbol $a \in \Sigma$ may be identified with a symbol for the set $\{a\}$. Consequently, it is enough to consider alphabets consisting of special symbols only. Since these special symbols denote subsets of the original alphabet, the set inclusion relation induces a partial order relation on the alphabet of the special symbols. This leads us to consider words over partially ordered alphabets. Based on the initial motivation, we will introduce a compatibility as a weaker form of equality, in a similar way as it is done for partial words. We will call words with such weak equality relation "fuzzy". Partial words thus become a special case of fuzzy words. Not all results on partial words admit straight translation to fuzzy words, since many of them are expressed in terms of the holes count in a partial word, while holes are specific for partial words only.

2 Basic notions

We denote $[n] = \{1, 2, \ldots, n\}$ for $n \geq 0$, by default $[0] = \varnothing$. Throughout this text $\Sigma = \{s_1.s_2, \ldots, s_k\}$ denotes a fixed partially ordered alphabet of size $k \geq 1$ with the partial order relation $\preccurlyeq$. A word w over Σ of length $n \geq 0$ with *symbol decomposition* $w = a_1 a_2 \ldots a_n$ is a mapping $w : [n] \to \Sigma$ such that $w(i) = a_i \in \Sigma$ for $i \in [n]$. The length of a word w is denoted as $|w|$. The empty word λ is the only word of length 0. If not stated otherwise, all words and languages are over Σ. The set of all words over Σ is denoted as Σ^*. A language over Σ is any subset of Σ^*. We will identify the singleton language $L = \{w\}$ with the word w a word of a unit length $w = a$, $a \in \Sigma$, with the symbol a. The *mirror image* of a word $w = a_1 a_2 \dots a_n$ is the word $w^R = a_n a_{n-1} \dots a_1$. The *concatenation* of two words $w_1 = a_1 a_1 \dots a_n$ and $w_2 = b_1 b_2 \dots b_m$ is the word $w_1 w_2 = a_1 a_2 \dots a_n b_1 b_2 \dots b_m$. If a word w can be expressed as $w = tuv$ then t, u, v are called *prefix, factor, suffix of* w, respectively (any of them may be empty). An ordered set $\iota = \{i_1 < \dots < i_m\} \subseteq [n]$, $m \geq 0$, is called *subword position* in w. The (*scattered*) *subword* occurring at position ι in w is the word $\sigma_w(\iota) = a_{i_1} a_{i_2} \dots a_{i_m}$, while ι is referred to as *occurrence* of $\sigma_w(\iota)$. We denote by $|w|_v$ the number of occurrences of the subword v in w. The only occurrence of λ in a word w is $\varnothing$, thus $|w|_\lambda = 1$. We will use the Kronecker δ notation: for two words $x, y \in \Sigma^*$ we denote $\delta_{x,y} =$ (if $x = y$ then 1 else 0). For words $w, v \in \Sigma^*$ and symbols $a, b \in \Sigma$ we have a rather straightforward equation (see (6.3.3) in [?])

$$|wa|_{vb} = |w|_{vb} + \delta_{a,b} |w|_v \,. \tag{1}$$

3 Fuzzy words

We extend the relation $\preccurlyeq$ to words from Σ^* as follows. For $x, y \in \Sigma^*$, $x \preccurlyeq y$ if $|x| = |y|$ and $x(i) \preccurlyeq y(i)$ for each $i \in [|x|]$. We will sometimes write $y \succcurlyeq x$ if $x \preccurlyeq y$. If the relation $\preccurlyeq$ on the set Σ^* is considered, we will refer to words from Σ^* as *fuzzy words* (though wee will mostly omit the adjective "fuzzy"). Let $x, y \in \Sigma^*$. We will say that x is *contained* in y if $x \preccurlyeq y$. Two words with symbol decompositions $x = a_0 a_1 \dots a_{m-1}$ and $y = b_0 b_1 \dots b_{n-1}$ are *fully compatible* (denoted as $x \Uparrow y$) if $m = n$ and, for each $0 \leq i < n$, a_i, either $a_i \preccurlyeq b_i$ or $b_i \preccurlyeq a_i$. Thus full compatibility of two words of length 1 means comparability (in the partial order relation) of the corresponding alphabet symbols. The words x, y are *compatible* (denoted as $x \uparrow y$) if there is a word $z \in \Sigma^*$ such that $z \preccurlyeq x$ and $z \preccurlyeq y$. Two fully compatible words x, y are compatible, since their the *greatest lower bound* $x \wedge y$ satisfies $x \wedge y \preccurlyeq x$ and $x \wedge y \preccurlyeq y$. Here $x \wedge y$ denotes the word of the same length as x and y, which contains at each position $i \in [|x|]$ the minimum of $x(i)$ and $y(i)$. We will call a fuzzy word *complete* if each symbol in its decomposition is minimal with respect to $\preccurlyeq$.

Proposition 1 *Let $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$ and let x, y be words of the same length. Then $x \backsim y$ iff for each $i \in [|x|]$, $x(i) \backsim y(i)$.*

Corollary 2 *Let $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$ and let x, y be complete words of the same length. Then $x \backsim y$ iff $x = y$.*

Some basic properties of the two relations are summarized in Propositions **??** and **??**. Since the relation $\succcurlyeq$ is itself a partial order relation, the assertions on $\preccurlyeq$ can be easily translated to similar assertions for $\succcurlyeq$. The same is true for the *least upper bound* $x \vee y$ of two fully compatible words x, y, being the word of the same length as x and y, which contains at each position $i \in [|x|]$ the maximum of $x(i)$ and $y(i)$.

Proposition 3 *Let u, v, x, y be words and let $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$.*

1. *If $x \Uparrow y$ then $x \uparrow y$.*
2. *If $x \preccurlyeq y$ or $y \preccurlyeq x$ then $x \Uparrow y$.*
3. *If $x \Uparrow y$ and $|x| = |y| = 1$ then $x \preccurlyeq y$ or $y \preccurlyeq x$.*
4. *Let $\backsim \neq \preccurlyeq$. Then $x \backsim y$ implies $y \backsim x$.*
5. *If $u \backsim v$ and $x \backsim y$ then $ux \backsim vy$.*
6. *If $ux \backsim vy$ and $|u| = |v|$ then $u \backsim v$ and $x \backsim y$.*
7. *If x, y are complete then $x \backsim y$ iff $x = y$.*

Proposition 4 *Let x, y, u, v be words, $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$. If $xu \backsim yv$ and $|x| \geq |y|$ then there are words z, t such that $x = zt$, $z \backsim y$, and $tu \backsim v$.*

For $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$, the $\backsim$-*closure* of a language L is the language $L^{\backsim} = \{y \in \Sigma^* | y \backsim x \text{ for some } x \in L\}$.

Partial words (more precisely their companions - as described in ([?])) may be considered as a special case of fuzzy words. In this case Σ contains the special symbol $\diamond$ and, for every symbol $a \in \Sigma$, $a \neq \diamond$, $a \preccurlyeq \diamond$, while no further non-trivial ordering relationship is valid. In this case strong compatibility and compatibility coincide.

4 The fuzzy closures and the Chomsky hierarchy

It is natural to investigate the relationship between the Chomsky type of a language and the Chomsky type of its closure. The basic relationship is described in the following theorem. For definitions of formal grammars, Chomsky hierarchy and related terms used in this section see [**?**].

Theorem 5 *Let $i \in \{0,1,2,3\}$ and $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$.*
1. For Σ a partially ordered alphabet Σ, if $L \subseteq \Sigma^*$ is a language of Chomsky type i then $L^{\backsim}$ is of type i.
2. There is a language L over a (totally) ordered alphabet $\Sigma = \{a \preccurlyeq b\}$, $a \neq b$, of type i, which (for $i \neq 3$) is not of type $i+1$, such that $L^{\backsim}$ is a regular language.

5 Periodicity and conjugacy

Periodicity of fuzzy words, as in the case of partial words, can be considered in a strong and a weak way. Let $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$. A *(strong) $\backsim$-period* of a word x is a positive integer p such that, for $1 \leq i, j \leq |u|$, $u\,(i) \backsim u\,(j)$ whenever $i \equiv j \bmod p$. A *weak $\backsim$-period* of a word x is a positive integer p such that $u\,(i) \backsim u\,(i+p)$ whenever $1 \leq i, i+p \leq |u|$. A word having a (strong)/ weak $\backsim$-period p is called *$p, \backsim$-periodic/weakly $p, \backsim$-periodic*, respectively. A word u is *primitive* if $v^i \preccurlyeq u$ implies $i = 1$ for each word v. Two words u, v are *$\backsim$-conjugate* if there are two words x, y such that $xy \backsim u$ and $yx \backsim v$. The following Proposition is implied by Corollary **??**.

Proposition 6 *Let $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$ and let u be a complete word with a strong or weak $\backsim$-period p. Then there are unique complete words x, y such that $|x+y| = p$, $|x| < p$, and $u = (xy)^n\, x$.*

Theorem 7 *Let $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$. Let u, v, z be words, $u \neq \lambda \neq v$. Then*
1. If $uz \backsim zv$ then uz, zv, and uzv are weakly $|u|\,, \backsim$-periodic.
2. If uzv is weakly $|u|\,, \backsim$-periodic then $uz \backsim zv$.
3. If uz, zv are weakly $|u|\,, \backsim$-periodic and $|z| \geq |u|$ then $uz \backsim zv$.
4. If $uz \backsim zv$ and there exists a complete word α, such that $\alpha \preccurlyeq uz$ and $\alpha \preccurlyeq zv$, and α is $|u|\,, \backsim$-periodic, then $xy \preccurlyeq u$, $yx \preccurlyeq v$ and $(xy)^n\, x \preccurlyeq z$ for some words x, y and $n \geq 0$.

6 Subword counts and extended Parikh matrices

The *Parikh mapping* ([**?**]) assigns to each word w the vector $[|w|_{s_1}, |w|_{s_2}, \ldots, |w|_{s_k}]$. Consider a word u with symbol decomposition $u = a_1 a_2 \cdots a_m, m \geq 1$. The *extended Parikh matrix mapping* ([**?**]) assigns to each word w the upper-triangular $(m+1) \times (m+1)$ matrix $\Psi_u(w)$ where the main diagonal consists of 1's and, for $1 \leq i \leq j \leq m$, the $(i, j+1)$-th element is $|w|_{a_i a_{i+1} \cdots a_j}$. Ψ_u is a morphism s mapping string concatenation to matrix product. In the case $u = s_1 s_2 \cdots s_k$ the mapping Ψ_u is the *Parikh matrix mapping* (originally introduced in [**?**]) denoted as Ψ_k and $\Psi_k(w)$ is the *Parikh matrix* of the word w. (Extended) Parikh matrices and are useful tools for investigation of subword occurrences in words.

We will consider different weaker forms of word occurrences, based on relations from $\{\uparrow, \Uparrow, \preccurlyeq\}$. We will extend the definition of the occurrence of a subword in a word to fuzzy words. Let $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$. We first extend the definition of our Kronecker-like symbol - we denote the extension as $\delta^{\backsim}$. For two words $x, y \in \Sigma^*$ we denote $\delta^{\backsim}_{x,y} =$ (if $x \backsim y$ then 1 else 0). Observe that, for complete words x, y, $\delta^{\backsim}_{x,y} = \delta_{x,y}$ and, for $\backsim \in \{\uparrow, \Uparrow\}$, $\delta^{\backsim}_{x,y} = \delta^{\backsim}_{y,x}$, which is generally not true for $\backsim = \preccurlyeq$. Let w be a word with symbol decomposition $w = a_0 a_1 \ldots a_{n-1}$ The (scattered) subword $\sigma_w(\iota)$ occurring at position ι in w is a *$\backsim$-occurrence* of a word v if $v \backsim \sigma_i\,(\iota)$. We denote by $|w|^{\backsim}_{v}$ the number of $\backsim$-occurrences of the subword v in w.

Example 8 *Let $\Sigma = \{a, b, c, d\}$ with the only non-trivial partial order relationships being $a \preccurlyeq d$, $b \preccurlyeq d$, $c \preccurlyeq d$. The word accdcbdab contains 6 $\Uparrow$-occurrences (being $\uparrow$-occurrences, as well) of the subword bdab:*

$\{3,4,6,8\}, \{3,4,7,8\}, \{3,5,6,8\}, \{3,5,7,8\}, \{3,6,7,8\}$ *and* $\{5,6,7,8\}$*, however , just* 2 $\preccurlyeq$*-occurrences of the word bdba:* $\{3,6,7,8\}$ *and* $\{5,6,7,8\}$ *. Thus* $|accdcbdab|_{bdab}^{\Uparrow} = 6$ *and* $|accdcbdab|_{bdab}^{\preccurlyeq} = 2$ *.*

The equality (**??**) can be extended to fuzzy words in the following way. Let $w, v \in \Sigma^*$, $a, b \in \Sigma$. Then

$$|wa|_{vb}^{\backsim} = |w|_{vb}^{\backsim} + \delta_{b,a}^{\backsim} |w|_{v}^{\backsim} . \tag{2}$$

Consider a word u with symbol decomposition $u = b_1 b_2 \cdots b_m$, $m \geq 1$. Denote, for $1 \leq i \leq j \leq m+1$, $u_{i,j} = b_i b_2 \cdots b_{j-1}$ (by default, $u_{i,i} = \lambda$) and $u_i = u_{1,i}$. Let $\mathcal{M}_p$ denote the set of all upper-triangular $p \times p$ matrices over real numbers, with the main diagonal consisting entirely of 1's. For $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$, we define a morphism $\Psi_u^{\backsim} : \Sigma^* \rightarrow \mathcal{M}_{m+1}$ as $[\Psi_u^{\backsim}(a)]_{i,j} = \delta_{u_j,u_i}^{\backsim} + \delta_{u_j,u_i a}^{\backsim}$, for $a \in \Sigma$, $1 \leq i, j \leq m+1$. Observe that, since the words u_i are of distinct length, $\delta_{u_i,u_j}^{\backsim} = \delta_{u_i,u_j}$ and $\delta_{u_j,u_i a}^{\backsim} = 1$ iff $j = i+1$ and $\delta_{b_j,a}^{\backsim} = 1$.

Theorem 9 *Let* $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$ *and* $w \in \Sigma^*$*. Then for each* $1 \leq i \leq j \leq m+1$, $[\Psi_u^{\backsim}(w)]_{i,j} = |w|_{u_{i,j}}^{\backsim}$*.*

In the remaining part of this section we will deal with formal power series. A *formal power series* (with integer coefficients) over Σ is a mapping $\mathbf{x} : \Sigma^* \rightarrow \mathbb{Z}$, where $\mathbb{Z}$ is the ring of integers. Following the usual conventions, we denote the value $\mathbf{x}(\alpha)$ as $\langle \mathbf{x}, \alpha \rangle$ and express $\mathbf{x}$ as $\sum_{v \in \Sigma^*} \langle \mathbf{x}, v \rangle v$. The *support* of $\mathbf{x}$ is the set $\{v \in \Sigma^* | \langle \mathbf{x}, v \rangle \neq 0\}$. The set of all power series over Σ, together with the sum and product operations defined as $\langle \mathbf{x} + \mathbf{y}, \alpha \rangle = \langle \mathbf{x}, \alpha \rangle + \langle \mathbf{y}, \alpha \rangle$, $\langle \mathbf{x} \cdot \mathbf{y}, \alpha \rangle = \sum_{uv=\alpha} \langle \mathbf{x}, u \rangle \langle \mathbf{y}, v \rangle$, respectively, forms a ring denoted as $\mathbb{Z}\langle\langle \Sigma \rangle\rangle$. Basic information on formal power series can be found in ([**?**]), as well as in ([**?**]). Our aim is to adapt the extended Parikh mapping from ([**?**]) to fuzzy words. We consider a finite factorial (i.e., containing with each word all its factors) language $L \subseteq \Sigma^*$ and the set $\mathbb{Z}_L$ consisting of all formal power series from $\mathbb{Z}\langle\langle \Sigma \rangle\rangle$ with support being a subset of L. It is proved in ([**?**]) that $\mathbb{Z}_L$ is a ring. The *L-projection* of a series $\mathbf{x} \in \mathbb{Z}\langle\langle \Sigma \rangle\rangle$ is the series $\pi_L(\mathbf{x}) = \sum_{v \in L} \langle \mathbf{x}, v \rangle v \in \mathbb{Z}_L$. It was proved in ([**?**]) that the extended Parikh morphism $\mathbf{\Pi}_L : \Sigma^* \rightarrow \mathbb{Z}_L$ defined as $\mathbf{\Pi}_L = \pi_L \circ \mu$, satisfies $\mathbf{\Pi}_L(w) = \sum_{v \in L} |w|_v v$ for any word w. Here $\mu : \Sigma^* \rightarrow \mathbb{Z}\langle\langle \Sigma \rangle\rangle$ is the Magnus (monoid) morphism defined, for $s_i \in \Sigma$, as $\mu(s_i) = 1 + s_i$. In a similar way, we define, for $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$, the fuzzy Magnus morphism $\mu^{\backsim} : \Sigma^* \rightarrow \mathbb{Z}\langle\langle \Sigma \rangle\rangle$ as $\mu^{\backsim}(s_i) = 1 + \sum_{a \backsim s_i} a$. Then the *extended fuzzy Parikh morphism* $\mathbf{\Pi}_L^{\backsim} : \Sigma^* \rightarrow \mathbb{Z}_L$ is defined as $\mathbf{\Pi}_L^{\backsim} = \pi_L \circ \mu^{\backsim}$.

Theorem 10 *Let* L *be a finite factorial language and* $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$*. Then, for each word* $w \in \Sigma^*$, $\mathbf{\Pi}_L^{\backsim}(w) = \sum_{v \in L} |w|_v^{\backsim} v$*.*

Again, as in ([**?**]), equality of the inverse Parikh series of a word (which always exists) and the so-called alternate series of the word's mirror image can be proved The *alternate series* of $\mathbf{x} \in \mathbb{Z}\langle\langle \Sigma \rangle\rangle$ is defined as $\overline{\mathbf{x}} = \sum_v (-1)^{|v|} \langle \mathbf{x}, v \rangle v$. It is easy to check that the mapping $\mathbf{x} \mapsto \overline{\mathbf{x}}$ is a ring morphism.

Theorem 11 *Let* $\backsim \in \{\uparrow, \Uparrow, \preccurlyeq\}$*. Let* L *be a factorial language, not containing any word bc, such that* $bc \backsim aa$ *for some symbol* $a \in \Sigma$*. Let* $w \in \Sigma^*$*. Then, in* $\mathbb{Z}_L$, $\mathbf{\Pi}_L(w)^{-1} = \overline{\mathbf{\Pi}_L(w^R)}$*.*

References

[1] F. Blanchet-Sadri, Algorithmic Combinatorics on Partial Words (Discrete Mathematics and Its Applications), Chapman & Hall/CRC, 2007.

[2] J. E. Hopcroft, R. Motwani, J. D. Ullman, Introduction to Automata Theory, Languages and Computation, 3rd ed., Pearson Addison-Wesley, Upper Saddle River, NJ, 2007.

[3] M. Lothaire, Combinatorics on words, Cambridge University Press, 1997.

[4] A. Mateescu, A. Salomaa, K. Salomaa, S. Yu, A sharpening of the Parikh mapping., RAIRO Theoretical Informatics and Applications 35 (6) (2001) 551–564.

[5] R. J. Parikh, On context-free languages, J. ACM 13 (4) (1966) 570–581.

[6] T.-F. Şerbănuţă, Extending Parikh matrices., Theor. Comput. Sci. 310 (1-3) (2004) 233–246.

[7] A. Černý, Generalizations of Parikh mappings, RAIRO-Theor. Inf. Appl. - to appear. URL `http://dx.doi.org/10.1051/ita/2009021`