Subword density of languages

M. Ito (伊藤 正美)
Faculty of Science
Kyoto Sangyo University
Kyoto 603, Japan

and

G. Paun
Institute of Mathematics
Str. Academiei 14
70109 Bucuresti, Romania

Let X be a nonempty finite set, called an *alphabet*. By X^* we denote the free monoid generated by X under the operation of catenation. Any subset of X^* is called a *language* over X and any element of X^* is called a *word* over X. The identity of X^* is denoted by 1, $X^* \setminus \{1\}$ is denoted by X^+, and $|x|$ is the length of $x \in X^*$. For $x \in X^*$, Sub(x) means the set $\{z \in X^* \mid$ there exist $u,v \in X^*$ such that $x = uzy\}$ (the set of subwords of x), and we can extend naturally this definition to languages:

$$L \subseteq X^*, \ \mathrm{Sub}(L) = \cup_{x\in L} \mathrm{Sub}(x).$$

For a given ordering of X, consider the lexicographic order on X^*; denote it by $\leq$, and write $x < y$ for $x \leq y, x \neq y$. For a set $M \subseteq X^*$, by min(M) we denote the minimum string in M, according to the lexicographic order.

In various contexts, related mainly to the theory of codes the following notion has been considered: a language $L \subseteq X^*$ is *dense* if Sub(L) = X^*. However, to separate languages in *dense* and *non-dense* seems to be rather rough; it is natural to look for degrees of density, at least for non-dense languages. This is the aim of the present note.

Namely, for a string $x \in X^*$, denote

$$\mathbf{H}_X(x) = \min\{X^* \setminus \mathrm{Sub}(x)\}$$

and $\mathbf{H}_X(x)$ is said to be the *height* of a word x.

Remark that $1 \in \mathrm{Sub}(x)$ and $x \in \mathrm{Sub}(x)$, hence in all cases $1 \neq \mathbf{H}_X(x)$.

For $L \subseteq X^*$, define

$$\mathbf{H}_X(L) = \{\mathbf{H}_X(x) \mid x \in L\}$$

and $\mathbf{H}_X(L)$ is said to be the *height* of a language L.

Intuitively, if we consider the strings of X^* arranged row by row according to their lengths and in lexicographic order on each row, and if we mark the subwords of x in this arrangement $\mathbf{H}_X(x)$ will be the smallest lexicographically from the shortest non-marked strings ("the front hole" in Sub(x)).

Example 1. $\mathbf{H}_{\{a,b\}}(ab) = aa$, $\mathbf{H}_{\{a,b\}}(a^ib^i) = ba$ for any $i \geq 2$ and hence $\mathbf{H}_{\{a,b\}}(\{a^nb^n \mid n \geq 1\}) = \{ab, ba\}$. □

Example 2. $\mathbf{H}_{\{a\}}(a^i) = a^{i+1}$ and $\mathbf{H}_{\{a\}}(L) = \{a\}L$ for any $L \subseteq a^*$, in particular, $\mathbf{H}_{\{a\}}(a^*) = a^+$. □

A similar result to Example 2 holds true for arbitrary alphabets.

Theorem 1. *For all* X, $\mathbf{H}_X(X^*) = X^+$. □

The following result follows from the proof of Theorem 1.

Theorem 2. *Let* $L \subseteq X^+$. *Then there exists* L_0 *such that* $\mathbf{H}_X(L_0) = L$. □

The following result shows relations between the density and height of languages.

Theorem 3. *A language $L \subseteq X^*$ is dense if and only if $\mathbf{H}_X(L)$ is infinite.* □

Theorem 4. *For each natural number n, there exists a dense language L_n such that $|w| > n$, for each string $w \in \mathbf{H}_X(L_n)$.* □

In the case of the one-letter apphabet, $|\mathbf{H}_{\{a\}}(a^n)| = n + 1$; for alphabets with at least two letters, $\mathbf{H}_X(x)$ is 'much" shorter than x.

Theorem 5. *Let $|X| \geq 2$. Then the following hold true:*

(i) For all k, $k \geq 1$ and all $x \in X^*$ with $|x| < |X|^k + k - 1$, $|\mathbf{H}_X(x)| \leq k$.

(ii) For all k, $k \geq 1$, there exists $y \in X^*$ such that $|y| = |X|^k + k - 1$ and $|\mathbf{H}_X(y)| = k + 1$.

(iii) For all k, $k \geq 1$, there exists $z \in X^*$ such that $|z| < |X|^k + k - 1$ and $|\mathbf{H}_X(z)| = k$. □

Theorem 3 provides some means to evaluate the density of a language $L \subseteq X^*$. Let $L, L' \subseteq X^*$ be two languages. If $\mathbf{H}_X(L)$ is infinite and $\mathbf{H}_X(L')$ is finite, we say that L is *more dense than* L'. Now we compare two languages $L, L' \subseteq X^*$ such that both $\mathbf{H}_X(L)$ and $\mathbf{H}_X(L')$ are finite.

Theorem 6. *Let $|X| \geq 2$ and let $L \subseteq X^*$. Moreover, let $\mathbf{H}_X(L)$ be finite. Then there exists $n \geq 1$ such that $\mathbf{H}_X{}^n(L) = \{a_1, a_2\}$ or $\mathbf{H}_X{}^n(L) = \{a_1\}$, where $X = \{a_1 < a_2 < \cdots < a_n\}$* and $\mathbf{H}_X{}^i(L) = \mathbf{H}_X(\mathbf{H}_X{}^{i-1}(L))$ for i, $i \geq 2$. □

Notation 1. Let $L \subseteq X^*$ be a language such that $\mathbf{H}_X(L)$ is finite. Then by $\rho(L)$ we denote the minimum number n satisfying the condition in Theorem 6.

Let $L, L' \subseteq X^*$ be two languages such that both $\mathbf{H}_X(L)$ and $\mathbf{H}_X(L')$ are finite. If $\rho(L) > \rho(L')$, we say that L is *more dense than*

L'. If $\rho(L) = \rho(L')$, then we say that L and L' have the *same density*.

Now we consider the case $\mathbf{H}_X(L)$ is infinite, i.e. L is dense. First, notice that it follows from Theorem 2 that $\mathbf{H}_X(L)$ is not necessarily dense when $L \subseteq X^*$ is dense. Therefore, we have the following two cases.

Case 1. There exists $n \geq 1$ such that $\mathbf{H}_X{}^n(L) = \{a_1, a_2\}$ or $\mathbf{H}_X{}^n(L) = \{a_1\}$. By $\sigma(L)$ we denote the minimum number n satisfying this condition.

Case 2. There is no such an $n \geq 1$. Then we denote $\sigma(L) = +\infty$.

Let $L, L' \subseteq X^*$ be two languages such that both $\mathbf{H}_X(L)$ and $\mathbf{H}_X(L')$ are infinite. If $\sigma(L) > \sigma(L')$, we say that L is *more dense than* L'. If $\sigma(L) = \sigma(L')$, then we say that L and L' have the *same density*. Thus we can classify the classe of languages from the point of view of the density of languages.

To conclude this note, we deal with the question whether the families in Chomsky hierarchy are closed under height operation.

Theorem 7. *The family of context-sensitive languages is not closed under height operation.* □

Theorem 8. *The family of context-sensitive languages is not closed under height operation.* □

The questions whether the famies of context-free languages and regular languages are closed under height operation are open.

References

1. Berstel and D. Perrin, Theory of Codes,

2. M. Ito and C.M. Reis, Left dense covers of semigroups, Proceedings of The Colloquium on Words, Languages and Combinatorics (1992) (World Scientific Publ. Co Pte Ltd, Singapore), to appear.

3. M. Ito and G. Tanaka, Dense property of initial literal shuffles, Inter. J. Computer Math. 34 (1990), 161 - 170.

4. A. Salomaa, Formal Languages, Academic Press, New York, 1973.

5. H.J. Shyr, Free Monoids and Languages, Lecture Notes, Institute of Appl. Math., National Chung-Hsing Univ., Taichung, Taiwan, 1991.