UNIVERSITY OF LIVERPOOL

LEVERHULME TRUST

The Leverhulme Research Centre for Functional Materials Design

(Un)ranking k-subsequence universal words

Duncan Adamson

15/06/2023

diamond Hartree Centre Imperial College London ICNS King Abdullah University of Science and Technology University of Southampton

Conventions

- All words are defined over the alphabet $\Sigma = [1, 2, \dots, \sigma]$. For simplicity, each symbol is also treated as its numeric value.
- Σ^n denotes the set of all words over Σ of length exactly n.
- Given a word w, the notation $w[i]$ is used to denote the i^{th} symbol of w.
- ε is used to denote the empty word.
- $A(w)$ is used to denote the alphabet formed by the symbols from w.

Subsequences (Just so we are all on the same page)

Definition

A **subsequence** of a word w is a sequence of that can be found be deleting some some set of symbols from w, i.e. a word that can be written as $w[i_1]w[i_2]\dots w[i_j]$ where $i_1 < i_2 < \dots < i_j$.

Definition

A **subword** of a word w is a contiguous subsequence of ww, i.e. a subsequence of the form $w[i]w[i+1]\dots w[i+j]$.

Example

123 *is a subsequence but not a subword of* **1**1**2**2**3**3. 1223 *is both a subsequence and a subword.*

Subsequences (Just so we are all on the same page)

Definition

A **subsequence** of a word w is a sequence of that can be found be deleting some some set of symbols from w, i.e. a word that can be written as $w[i_1]w[i_2]\dots w[i_j]$ where $i_1 < i_2 < \dots < i_j$.

Definition

A **subword** of a word w is a contiguous subsequence of ww, i.e. a subsequence of the form $w[i]w[i+1]\dots w[i+j]$.

Example

123 *is a subsequence but not a subword of* 1**1223**3*.* 1223 *is both a subsequence and a subword.*

k-Subsequence Universality

Definition

A word w is k-subsequence universal over the alphabet Σ if and only if every word in Σ^k is a subsequence of w. The set of all k-subsequence universal words of length n is denoted $\mathcal{U}(n, k, \sigma)$.

k-Subsequence Universality Example

Example

Let $w = 11223231$. *Then* w *is* 2 *subsequence universal over* $\Sigma = [1, 2, 3]$.

11	**11**223231
12	**1**1**2**23231
13	**1**122**3**231
21	11**2**2323**1**
22	11**22**3231
23	11**2**2**3**231
31	1122**3**23**1**
32	1122**32**31
33	1122**3**2**3**1

k-Subsequence Universality Example

Example

Let $W = 22323111$. *Then* w *is* **not** 2 *subsequence universal over* $\Sigma = [1, 2, 3]$.

11	22323**11**1
12	22323**1**11 (2)
13	22323**1**11 (3)
21	**2**2323**1**11
22	**22**323111
23	**2**2**3**23111
31	22**3**23**1**11
32	22**32**3111
33	22**3**2**3**111

Universality Index

Definition

The **universality index** of word w, denoted $\zeta(w)$, is the maximum value such that w is $\zeta(w)$ universal.

Example

The universality index of 1122321 *is* $\zeta(11223231) = 2$, *the universality index of* 22323111 *is* $\zeta(22323111) = 1$.

Combinatorial Results

Arches

Definition

An **Arch** in a word w is a subword $w[i]w[i+1]\dots w[i+j]$ containing each symbol in Σ at least once, and the symbol $w[i+j]$ **exactly** once.

Arches

Example

Given the word $w = 11231123$ *the possible arches are:*

- $\mathbf{1123}1123$
- $1\mathbf{123}1123$
- $11\mathbf{231}123$
- $112\mathbf{3112}3$
- $1123\mathbf{1123}$
- $11231\mathbf{123}$

Universal Subsequences and Free Symbols in Arches

Definition

Given an arch w, the **Universal Subsequence** of w is the subsequence u of length σ such that $u[1]$ is the first symbol to appear in w, $u[2]$ is the second unique symbol, and $u[i]$ is the i^{th} unique symbol.

Definition

Given an arch w and index i, $w[i]$ is a **Free Symbol** if and only if there exsits some index $j < i$ such that $w[i] = w[j]$.

12112233214

Arch Factorisations

Definition

Given the word w, the **Arch Factorisation** of w, denoted $Arch(w)$ is the set of words $Arch(w) = u_1, u_2, \ldots, u_m, v$ such that:

- $w = u_1 u_2 \ldots u_m v$,
- $\forall i \in [m], u_i$ is an Arch,
- v is not an arch.

Example

Given the word $w = 112322133211$, $Arch(w) = 1123, 2213, 321, 1$.

Arch Factorisations and Universality

Theorem (Day et al.[1])

A word $w \in \Sigma^n$ is k-subsequence universal over Σ if and only if $Arch(w)$ contains at least k arches. Further, $Arch(w)$ can be computed in $O(n)$ time.

[1] Joel D. Day et al. "The Edit Distance to k-Subsequence Universality". In: *38th International Symposium on Theoretical Aspects of Computer Science (STACS 2021)*. Ed. by Markus Bläser et al. Vol. 187. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 25:1–25:19.

Counting k-subsequence universal words

High Level Sketch

- We introduce the set $S(v, n)$,containing every k-subsequence universal word in Σ^n with the prefix v, formally

$$S(v, n) = \{vu \mid u \in \Sigma^{n-|v|}, \zeta(vu) \geq k\}.$$

- Note that $S(\varepsilon, n)$ is the set if all k-subsequence universal words of length n.
- **Idea:** use the size of $S(vx, n)$ to count the size of $S(v, n)$, for every $x \in \Sigma$.

Using $S(v, n)$

Observation

Given $v \in \Sigma^{\ell}, S(v, n) = \bigcup_{x \in \Sigma} S(vx, n)$ *and further, for any pair of symbols* $x, y \in \Sigma$ *such that* $x \neq y$, $S(vx, n) \cap S(vy, n) = \emptyset$.

Counting the size of $S(v, n)$

Lemma

Given the word v with the arch decomposition $Arch(v) = u_1, u_2, \ldots, u_m, v'$. Then, given the pair of symbols $x, y \in \Sigma$ such that both x and y are in v', the size of $S(vx, n)$ is the same as $S(vy, n)$.

Proof (Sketch).

Let w be a word such that $\zeta(vxw) = k$ with the arch decomposition $Arch(vxw) = w_1, w_2, \ldots, w_k w'$. Note that $w_{m+1} = v'xu$, for some prefix u of w such that u contains every symbol in Σ that does not appear in $v'x$, and by extension v'. Therefore $v'yu$ is an arch and hence $\zeta(vyw) = k$. □

Counting the size of $S(v, n)$

Lemma

Given the word v with the arch decomposition $Arch(v) = u_1, u_2, \ldots, u_m, v'$. Then, given the pair of symbols $x, y \in \Sigma$ such that neither x nor y are in v', the size of $S(vx, n)$ is the same as $S(vy, n)$.

Proof (Sketch).

Let w be a word such that $\zeta(vxw) = k$ with the arch decomposition $Arch(vxw) \geq w_1, w_2, \ldots, w_k w'$. Note that $w_{m+1} = v'xu$, for some prefix u of w such that u contains every symbol in Σ that does not appear in $v'x$, and by extension v'. Now let u' be the word constructed by substituting every occurrence of y in u with x, and every occurrence of x in u with y. Then $v'yu'$ is an arch and hence $\zeta(vyw) \geq k$. □

Counting the size of $S(v, n)$

- Using these observations, the size of $S(v, n)$, where $Arch(v) = u_1, u_2, \dots, u_m v'$, can be computed by splitting it in to two cases:
  - The size of the set $S(vx, n)$, where x is some symbol in v'.
  - The size of the set $S(vy, n)$, where y is some symbol not in v'.

  Combining these gives the equation:

$$\mid S(v, n) \mid = \mid A(v') \mid\mid S(vx, n) \mid + (\sigma - \mid A(v') \mid) \mid S(vy, n) \mid .$$

Recursively Counting $S(v, n)$

- Using the outline above, we make a new function $CS(q, m, c)$.
- Given some prefix $v \in \Sigma^*$ such that $Arch(v) = v_1, v_2, \dots, v_\ell, v'$, to count the size of the set $S(v, n)$, the parameters for $CS(q, m, c)$ are derived as follows:
  - q is equal to the number of symbols in Σ that are not in v', $\sigma- \mid A(v') \mid$.
  - c is the (minimum) number of Arches that need to be present in each suffix in $S(v, n)$, i.e. $k - \ell$.
  - m is the remaining number of "free" symbols (symbols that do not need to belong to any arch), i.e. $n - (\mid v \mid + q + (c - 1)\sigma)$.

$CS(q, m, c)$

Using the same two cases as before, the value of $CS(q, m, c)$ is split in to two main cases:

- Counting the size of the set $S(vx, n)$, where x is some symbol in v', equal to $CS(q, m-1, c)$ as any such x must be a free symbol, i.e. not in the universal subsequence of the arch containing it. Further, there are $(\sigma - q)$ possible values of x.
- Counting the size of the set $S(vy, n)$, where y is some symbol not in v', equal to $CS(q-1, m, c)$m as any such symbol must be one of the q symbols that do not appear in v'. Further, there are q possible values of y.

$CS(q, m, c)$

Additionally, there are a set of three special cases:

- If $q = 0$ and $c > 0$, then $v' = \varepsilon$, and whatever the next symbol is, it must belong to the universal subsequence of the first arch of the suffix, giving the size of $S(vx, n)$ as 0 and $S(vy, n)$ as $CS(\sigma - 1, m, c - 1)$. Note that there are σ possible values of y.
- If $c = 0$ and $q = 0$, then every remaining symbol is "free" in that it does not matter if there are any more arches. Therefore, the size of $S(v, n)$ is σ^m.
- If $m = 0$ then every remaining symbol must be in the universal subsequence of some arch, giving $\mid S(v, n) \mid = q!(\sigma!)^c$.

$CS(q, m, c)$

$$CS(q,m,c) = \begin{cases} \sigma^m & q = 0, c = 0 \\ q!(\sigma!)^c & m = 0 \\ \sigma CS(\sigma - 1, m, c - 1) & q = 0, c > 0 \\ (\sigma - q)CS(q, m - 1, c) + qCS(q - 1, m, c) & q > 0, c > 0, m > 0 \end{cases}$$

Counting the number of k-subsequence universal words

Theorem

The size of $\mathcal{U}(n, k, \sigma)$ can be computed in $O(nk\sigma)$ time.

Ranking

Where we actually talk about the title of the paper

Ranking

Definition

Let $\mathcal{U}(n, k, \sigma)$ be the set of all k-subsequence universal words of length n over the alphabet $[1, 2, \dots, \sigma]$. The rank of some word $w \in \mathcal{U}(n, k, \sigma)$ is the number of words in $\mathcal{U}(n, k, \sigma)$ that are lexicographically smaller than w.

High Level Idea

- Starting with the empty word ε, the idea is to count the number of words smaller than the input word w, sharing a a given prefix of w.
- First, we count the number of words starting with any symbol $x < w[1]$, given by $(w[1] - 1)CS(\sigma - 1, n - k\sigma, k)$.
- Then, we count the number of words with the prefix $w[1]$ followed by some symbol $x < w[2]$. This is split in to two cases. If $x = w[1]$, then the number of such words is $CS(q - 1, m - 1, k)$, otherwise the number of such words is $CS(q - 2, m, k)$.

High Level Idea

At the i^{th} step, we count the number of words with the prefix $w[1]w[2]\dots w[i]$ followed by some $x < w[i+1]$. Letting $Arch(w[1]w[2]\dots w[i]) = v_1 v_2 \dots v_\ell w'$, $q = \sigma - A(w')$, and $m = n - (i + q + (k - \ell - 1)\sigma)$ the number of such words is given by:

$$\sum_{x \in \Sigma} \begin{cases} 0 & x \geq w[i+1] \\ CS(q-1, m, k-\ell-1) & x \notin A(w') \\ CS(q, m-1, k-\ell-1) & x \in A(w') \end{cases}$$

Full Ranking Algorithm

$$\sum_{i \in [1 \dots n]} \sum_{x \in \Sigma} \begin{cases} 0 & x \geq w[i] \\ CS(q-1, m, k-\ell-1) & x \notin A(w') \\ CS(q, m-1, k-\ell-1) & x \in A(w') \end{cases}$$

Ranking Efficiently

- Our counting proccess works by computing the value of $CS(q, m, c)$, for every $q \in [1, 2, \ldots, \sigma], m \in [1, 2, \ldots, n - k\sigma]$ and $c \in [1, 2, \ldots, k]$ in $O(nk\sigma)$ time. Therefore, we assume this has been precomputed.
- At each step, the algorithm needs to find the value of $CS(q, m, c)$ for at most σ-values.
- As there are n such steps, this requires the table of $CS(q, m, c)$ values at most $O(n\sigma)$ times.

Ranking Result

Theorem

The rank of a word w within the set $\mathcal{U}(n, k, \sigma)$ can be computed in $O(nk\sigma)$ time.

Unranking and Enumeration

Unranking

Definition

Let $\mathcal{U}(n, k, \sigma)$ be the set of all k-subsequence universal words of length n over the alphabet $[1, 2, \ldots, \sigma]$. The unranking problem asks, for a given input value i, what is the word in $\mathcal{U}(n, k, \sigma)$ with a rank of i.

Unranking the j^{th} symbol

x satisfies:

$$\sum_{y \in [1,2,\ldots,x-1]} | S(w[1]w[2]\ldots w[j-1]y) | < i$$

$$\sum_{y \in [1,2,\ldots,x]} | S(w[1]w[2]\ldots w[j-1]y) | \geq i.$$

Outline

- Letting w be the word with a rank of i, the value of $w[1]$ is determined by finding the symbol x such that $xCS(q-1, m, k) \leq i < (x+1)CS(q-1, m, k)$.
- Proceeding iteratively, the value of $w[j]$ is determined by finding the symbol x such that:
  - The rank r_s of the word v_s, defined as the smallest word in $\mathcal{U}(n, k, \sigma)$ with the prefix $w[1]w[2]\ldots w[j-1]x$, is less than or equal to i.
  - The rank r_l of the word v_l, defined as the largest word in $\mathcal{U}(n, k, \sigma)$ with the prefix $w[1]w[2]\ldots w[j-1]x$, is greater than or equal to i.

Theorem

The i^{th} word in the set $\mathcal{U}(n, k, \sigma)$ can be computed in $O(n\sigma)$ time after $O(nk\sigma)$ preproccessing.

Enumerating

Theorem

Every word in $\mathcal{U}(n, k, \sigma)$ can be output with at most $O(n\sigma)$ delay after $O(nk\sigma)$ preproccessing time.

Conclusion

- An $O(nk\sigma)$ time algorithm for counting the size of $\mathcal{U}(n, k, \sigma)$;
- An $O(nk\sigma)$ time algorithm for ranking words in the set $\mathcal{U}(n, k, \sigma)$;
- An $O(nk\sigma)$ time algorithm for unranking words from the set $\mathcal{U}(n, k, \sigma)$;
- An algorithm for enumerating the set $\mathcal{U}(n, k, \sigma)$ with $O(n\sigma)$ delay after $O(nk\sigma)$ preprocessing.

Future Work

- Finding a better way of counting the number of k-subsequence universal words.
  - As well as being an interesting result on its own, this may allow us to speed up the ranking, unranking and enumerating results.
- Reduce the delay in the enumeration proccess.
  - This should either be to $O(n)$, if each word is explicitly represented, or sub-linear if the word in the memory is simply being updated at each time step.