Generating the k-neighbourhood of sequences using alignments

France Paquet-Nadeau
Supervisor: Cédric Chauve, Co-supervisor: Marni Mishna

Simon Fraser University

December 1, 2016

The Problem

Given a word P of length n, an alphabet Σ and a maximal distance k, what is the k-neighbourhood of P? How can we adapt the current method to generate it exactly?

Definitions

- An **alphabet** Σ is a set of letters.
- We use the letters of the alphabet to create words.

Example

$\Sigma = \{a, b, c\}$
possible words: ab, bbb, $acba$, $abaa$, $cbabcbabcba$

Formally, words are called **sequences**.

An **alignment** between two sequences

a	b	a	a
a	-	a	c

is a **supersequence**.

The allowed operations for supersequences are:

$\begin{array}{|c|}\hline a \\ \hline a \\ \hline\end{array}$ match, $\begin{array}{|c|}\hline a \\ \hline x \\ \hline\end{array}$ substitution, $\begin{array}{|c|}\hline - \\ \hline a \\ \hline\end{array}$ insertion and $\begin{array}{|c|}\hline a \\ \hline - \\ \hline\end{array}$ deletion.

The Method

We use the previous operations to create alignments that generate the words of the neighbourhood. We use recurrence relations to generate the desired words. The bottom line of an alignment, once we remove the gaps, gives us a word from the neighbourhood.

a	b	a	a
a	-	a	c

$\Rightarrow a - ac \Rightarrow aac$

The Method

We use the previous operations to create alignments that generate the words of the neighbourhood. We use recurrence relations to generate the desired words. The bottom line of an alignment, once we remove the gaps, gives us a word from the neighbourhood.

a	b	a	a
a	-	a	c

$\Rightarrow a - ac \Rightarrow aac$

The current upper bound is on the number of words in the condensed-neighborhood.

Example

aba is a prefix of *abaa*, *abacb* and *ababbbac*

Recurrences

Let $S(k, d)$ represent the number edit scripts with d differences in the remaining k symbols of the sequence.

Lemma

If $k \leq d$ or $d = 0$ then $S(k, d) = 1$. Otherwise,

$$\begin{aligned} S(k,d) = {} & S(k-1,d) + (s-1)S(k-1,d-1) \\ & + (s-1)\sum_{j=0}^{d-1} s^j S(k-2, d-1-j)) \\ & + (s-1)^2 \sum_{j=0}^{d-2} s^j S(k-2, d-2-j) \\ & + \sum_{j=0}^{d-1} S(k-2-j, d-1-j) \end{aligned}$$

Proof

The bound on the size of the condensed-neighbourhood is given by:

$$\chi_d(k) \leq S(k,d) + \sum_{j=1}^{d} s^j S(k-1, d-j)$$

Proof

The bound on the size of the condensed-neighbourhood is given by:

$$\chi_d(k) \leq S(k,d) + \sum_{j=1}^{d} s^j S(k-1, d-j)$$

In order to reduce the redundancy of the words generated, we only authorize the following combinations of operations:

1. a match followed by any operation
2. a substitution followed by any operation
3. insertion followed by other insertions and terminate with a match
4. deletion followed by other deletions and terminate with a match

Suppose we have a query of length n:

n	n-1	...	k+1	k	k-1	k-2	...	2	1

Suppose we have a query of length n:

n	n-1	...	k+1	k	k-1	k-2	...	2	1

In terms of $S(k, d)$:

1. **match followed by any operation**

...	k	k-1	...	1
...	a	x	...	x
...	a	$\vdash$	d differences	$\dashv$

Suppose we have a query of length n:

n	n-1	...	k+1	k	k-1	k-2	...	2	1

In terms of $S(k, d)$:

1. **match followed by any operation**

...	k	k-1	...	1
...	a	x	...	x
...	a	$\vdash$	d differences	$\dashv$

$\rightarrow S(k-1, d)$

Suppose we have a query of length n:

n	n-1	...	k+1	k	k-1	k-2	...	2	1

In terms of $S(k, d)$:

1. **match followed by any operation**

...	k	k-1	...	1
...	a	x	...	x
...	a	$\vdash$	d differences	$\dashv$

$\rightarrow S(k-1, d)$

2. **substitution followed by any operation**

Suppose we have a query of length n:

n	n-1	...	k+1	k	k-1	k-2	...	2	1

In terms of $S(k, d)$:

1. **match followed by any operation**

...	k	k-1	...	1
...	a	x	...	x
...	a	$\vdash$	d differences	$\dashv$

$\rightarrow S(k-1, d)$

2. **substitution followed by any operation**

   1. In the case where the next operation is NOT an insertion:

...	k	k-1	...	1
...	a	x	...	x
...	y	$\vdash$	$d-1$ differences	$\dashv$

$\rightarrow (s-1)S(k-1, d-1)$

Suppose we have a query of length n:

n	n-1	...	k+1	k	k-1	k-2	...	2	1

In terms of $S(k, d)$:

1. **match followed by any operation**

...	k	k-1	...	1
...	a	x	...	x
...	a	$\vdash$	d differences	$\dashv$

$\rightarrow S(k-1, d)$

2. **substitution followed by any operation**
   1. In the case where the next operation is NOT an insertion:

...	k	k-1	...	1
...	a	x	...	x
...	y	$\vdash$	$d-1$ differences	$\dashv$

$\rightarrow (s-1)S(k-1, d-1)$

   2. The next operation is an insertion
      $\rightarrow (s-1)I(k-1, d-2)$

3. **sequence of insertions that terminate with a match**

3. **sequence of insertions that terminate with a match**
   $I(k, d)$ is the number of d edit scripts that immediately follow one or more insertions after the $k+1$ symbol of the query.

3. **sequence of insertions that terminate with a match**

$I(k,d)$ is the number of d edit scripts that immediately follow one or more insertions after the $k+1$ symbol of the query.

$$I(k,d) = sI(k,d-1) + S(k-1,d)$$

...	k+1	-	k	k-1	...	1
		insert	$\vdash$	d differences		$\dashv$

$=$

...	k+1	-	-	k	k-1	...	1
		insert	insert	$\vdash$	$d-1$ differences		$\dashv$

$+$

...	k+1	-	k	k-1	k-2	...	1
		insert	match	$\vdash$	d differences		$\dashv$

3. **sequence of insertions that terminate with a match**

$I(k,d)$ is the number of d edit scripts that immediately follow one or more insertions after the $k+1$ symbol of the query.

$$I(k,d) = sI(k,d-1) + S(k-1,d)$$

...	k+1	-	k	k-1	...	1
		insert	$\vdash$	d differences		$\dashv$

$=$

...	k+1	-	-	k	k-1	...	1
		insert	insert	$\vdash$	$d-1$ differences		$\dashv$

$+$

...	k+1	-	k	k-1	k-2	...	1
		insert	match	$\vdash$	d differences		$\dashv$

$$\rightarrow I(k-1,d-1) = \sum_{j=0}^{d-1} s^j S(k-2,d-1-j)$$

$$\rightarrow I(k-1,d-2) = \sum_{j=0}^{d-2} s^j S(k-2,d-2-j)$$

4. **sequence of deletions that terminate with a match**
   $D(k, d)$ is the number of d edit scripts that immediately follow a deletion of the $k+1$ symbol.

4. **sequence of deletions that terminate with a match**
$D(k,d)$ is the number of d edit scripts that immediately follow a deletion of the $k+1$ symbol.

$$D(k,d) = D(k-1,d-1) + S(k-1,d)$$

$$\boxed{\begin{array}{cccccc} \ldots & \text{k+1} & \text{k} & \text{k-1} & \ldots & 1 \\ & \text{delete} & \vdash & & d \text{ differences} & \dashv \end{array}} =$$

$$\boxed{\begin{array}{cccccc} \ldots & \text{k+1} & \text{k} & \text{k-1} & \ldots & 1 \\ & \text{delete} & \text{delete} & \vdash & d-1 \text{ differences} & \dashv \end{array}}$$

$$+$$

$$\boxed{\begin{array}{cccccc} \ldots & \text{k+1} & \text{k} & \text{k-1} & \ldots & 1 \\ & \text{delete} & \text{match} & \vdash & d \text{ differences} & \dashv \end{array}}$$

4. **sequence of deletions that terminate with a match**
$D(k,d)$ is the number of d edit scripts that immediately follow a deletion of the $k+1$ symbol.

$$D(k,d) = D(k-1,d-1) + S(k-1,d)$$

$$\boxed{\begin{array}{cccccc} \ldots & \text{k+1} & \text{k} & \text{k-1} & \ldots & 1 \\ & \text{delete} & \vdash & & d \text{ differences} & \dashv \end{array}} =$$

$$\boxed{\begin{array}{ccccccc} \ldots & \text{k+1} & \text{k} & \text{k-1} & & \ldots & 1 \\ & \text{delete} & \text{delete} & \vdash & & d-1 \text{ differences} & \dashv \end{array}}$$

$$+$$

$$\boxed{\begin{array}{ccccccc} \ldots & \text{k+1} & \text{k} & \text{k-1} & & \ldots & 1 \\ & \text{delete} & \text{match} & \vdash & & d \text{ differences} & \dashv \end{array}}$$

$$\rightarrow D(k-1,d-1) = \sum_{j=0}^{d-1} S(k-2-j, d-1-j)$$

1. a match followed by any operation: $S(k-1,d)$
2. a substitution followed by:
   - anything but insertion: $(s-1)S(k-1,d-1)$
   - insertion:
$$(s-1)^2\sum_{j=0}^{d-2} s^j S(k-2,d-2-j)$$
3. insertion followed by other insertions and terminate with a match:
$$(s-1)\sum_{j=0}^{d-1} s^j S(k-2,d-1-j)$$
4. deletion followed by other deletions and terminate with a match:
$$\sum_{j=0}^{d-1} S(k-2-j,d-1-j)$$

Recurrences

Let $S(k,d)$ represent the number edit scripts with d differences in the remaining k symbols of the sequence.

Lemma

If $k \le d$ or $d = 0$ then $S(k,d) = 1$. Otherwise,

$$\begin{aligned} S(k,d) = S(k-1,d) &+ (s-1)S(k-1,d-1) \\ &+ (s-1)\sum_{j=0}^{d-1} s^j S(k-2,d-1-j)) \\ &+ (s-1)^2 \sum_{j=0}^{d-2} s^j S(k-2,d-2-j) \\ &+ \sum_{j=0}^{d-1} S(k-2-j,d-1-j) \end{aligned}$$

Ongoing Research

- With the recurrences coded, we can see which words are repeated and how many times.

Ongoing Research

- With the recurrences coded, we can see which words are repeated and how many times.

Example

$P = abaa$, $d = 2$, $\Sigma = \{a, b\}$
Total number of words produced: 49
Number of words generated more than once: 12

Ongoing Research

- With the recurrences coded, we can see which words are repeated and how many times.

Example

$P = abaa$, $d = 2$, $\Sigma = \{a, b\}$
Total number of words produced: 49
Number of words generated more than once: 12

- Need to identify patterns that cause the redundancy in order to improve the bound.

Ongoing Research

- With the recurrences coded, we can see which words are repeated and how many times.

Example

$P = abaa$, $d = 2$, $\Sigma = \{a, b\}$
Total number of words produced: 49
Number of words generated more than once: 12

- Need to identify patterns that cause the redundancy in order to improve the bound.

Example

The word *aba* can be generated in two different ways.

a	b	a	a
a	b	-	a

and

a	b	a	a
a	b	a	-