# The Optimal Copypasta Problem

avaxzat

July 16, 2016

## 1 Introduction

### 1.1 The Copypasta Problem

Let $\Sigma$ be a finite alphabet. A *context* is a tuple $(b, c, s)$ where $b$, $c$ and $s$ are strings over $\Sigma$, respectively called the *output buffer*, *clipboard* and *selection buffer*. We let $\epsilon$ denote the empty string. A *copypasta* is a string of the form $P_1; \ldots; P_n$ where the $P_k$ are any of $\mathsf{Write}(\sigma, i)$, $\mathsf{Select}(i, j)$, $\mathsf{Copy}$ or $\mathsf{Paste}(i)$, for all $\sigma \in \Sigma$ and integers $i, j$. Informally, these statements have the following meaning:

- $\mathsf{Write}(\sigma, i)$. Write symbol $\sigma$ to position $i$ in the output buffer.
- $\mathsf{Select}(i, j)$. Select the range of characters from index $i$ up to and including index $j$ in the output buffer and copy them to the selection buffer, overwriting its previous contents.
- $\mathsf{Copy}$. Copy the selection buffer to the clipboard.
- $\mathsf{Paste}(i)$. Insert the string in the clipboard at index $i$ in the output buffer.

The optimal copypasta problem can now be stated as follows:

> Given a string $t$ over $\Sigma$. Find the smallest copypasta $P$ such that $(\epsilon, \epsilon, \epsilon) \vdash P$ rewrites to $(t, c, s)$ according to the rules of Figure 1, where $c$ and $s$ are arbitrary strings over $\Sigma$.

### 1.2 Basic properties

**Proposition 1.1.** *Let $t$ be a string over $\Sigma$ containing a total of $n$ unique characters. The length of the optimal copypasta for $t$ will be at least $n$.*

*Proof.* Every unique character in $t$ needs to be written to the output buffer at least once, so any copypasta for $t$ must contain at least $n$ $\mathsf{Write}$ statements. □

**Proposition 1.2.** *The optimal copypasta for $t$ has length at most $|t|$.*

*Proof.* This is exactly the length of the copypasta that explicitly writes out $t$ using $\mathsf{Write}$ statements. □

## 2 A simple algorithm

Algorithm 1 shows the pseudocode.

$$\frac{(b, c, \epsilon) \vdash \mathsf{Write}(\sigma, i)}{(b_1 \dots b_{i-1} \sigma b_i \dots b_n, c, \epsilon)} \text{ E-WRITE}$$

$$\frac{(b, c, s) \vdash \mathsf{Select}(i, j)}{(b, c, b_i \dots b_j)} \text{ E-SELECT}$$

$$\frac{(b, c, s) \vdash \mathsf{Copy}}{(b, s, s)} \text{ E-COPY}$$

$$\frac{(b, c, s) \vdash \mathsf{Paste}(i)}{(b_1 \dots b_{i-1} c b_i \dots b_n, c, s)} \text{ E-PASTE}$$

$$\frac{\dfrac{(b, c, s) \vdash A}{(b', c', s')} \qquad (b, c, s) \vdash A; B}{(b', c', s') \vdash B} \text{ E-SEQ}$$

Figure 1: Rewrite rules for the copypasta problem

**Data:** a string $t$, $|t| = n$
**Result:** a copypasta for $t$

```
i ← 1;
while i ≤ n do
    if t[1 : i − 1] and t[i : n] have a common substring of length at least 3 which repeats at index i then
        Let s be the longest common substring of t[1 : i − 1] and t[i : n] which repeats at index i.
        Let j be the index of the first occurence of s in t[1 : i − 1].
        Output Select(j, j + |s| − 1); Copy; Paste(i).
        i ← i + |s|;
    else
        Output Write(t_i, i).
        i ← i + 1;
    end
end
```

**Algorithm 1:** A simple solver for the copypasta problem