A CFL that has a LARGE CFG but a small CSL

Exposition by William Gasarch

1 Introduction

In this section we give an example, due to Ellul et al [EKSW05], of a CFL L_n whose CFG has to be large, but whose CSG is small.

Def 1.1 A CFG $G = (N, \Sigma, S, R)$ is in *Chomsky Normal Form* if every rule in R is either of the form $A \rightarrow BC$ or $A \rightarrow \sigma$ where $A, B, C \in N$ aand $\sigma \in \Sigma$.

The following definition is not standard but it will help us standardize things.

Def 1.2 A CSG $G = (N, \Sigma, S, R)$ is in *Chomsky Normal Form* if every rule in R is either of the form $A \rightarrow CD$ OR $AB \rightarrow CD$ OR $A \rightarrow \sigma$ where $A, B, C, D \in N$ aand $\sigma \in \Sigma$.

In this manuscript e will assume that CFG's and CSL's are in Chomsky Normal Form and we will measure the size of a grammar by the number of nonterminals.

2 The Proof

Def 2.1 If F is a finite set then $PERM(F)$ is the set of all permuations of elements of F. Note that $PERM(F)$ has $|F|!$ elemenents.

Lemma 2.2 *Let* $0 < \beta < 1$. *Then* $\frac{n!}{(\beta n)!((1-\beta)n)!} = \Theta\left(\frac{1}{\sqrt{n}}\left(\frac{1}{(1-\beta)^{1-\beta}\beta^{\beta}}\right)^n\right)$

Proof:

By Stirling's Formula $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$. We use this in the form $n! = \Theta(\sqrt{n}(\frac{n}{e})^n)$. We omit the symbol Θ in our calculations.

$$(\beta n)!(((1-\beta)n))! \sim \sqrt{\beta n}\left(\frac{\beta n}{e}\right)^{\beta n}\sqrt{((1-\beta)n)}\left(\frac{(1-\beta n)}{e}\right)^{(1-\beta)n} =$$

$$\frac{(\sqrt{\beta(1-\beta)})n}{e^n}(\beta n)^{\beta n}((1-\beta)n)^{(1-\beta)n} = \frac{(\sqrt{\beta(1-\beta)})n}{e^n}((1-\beta)n)^n\left(\frac{\beta}{1-\beta}\right)^{\beta n}$$

Inverting this and multiplying by $\sqrt{n}(\frac{n}{e})^n$ yields

$$\sqrt{n}\left(\frac{n}{e}\right)^n \frac{e^n}{\sqrt{\beta(1-\beta)n}}\frac{1}{((1-\beta)n)^n}\left(\frac{1-\beta}{\beta}\right)^{\beta n} = \frac{1}{\sqrt{n}}\frac{1}{(1-\beta)^n}\left(\frac{1-\beta}{\beta}\right)^{\beta n} =$$

$$\frac{1}{\sqrt{n}}\left(\frac{(1-\beta)^{\beta-1}}{\beta^\beta}\right)^n = \frac{1}{\sqrt{n}}\left(\frac{1}{(1-\beta)^{1-\beta}\beta^\beta}\right)^n$$

∎

Def 2.3 If $n \in \mathsf{N}$ then $[n] = \{1, \ldots, n\}$

Theorem 2.4 *For all n there exists a language L_n such that*

1. *Any Chomsky Normal Form CFG for L_n requires $\Omega\left(\frac{1.89^n}{n^{3/2}}\right)$ nonterminals.*
2. *There is a CSL for L_n that has $O(n^2)$ nonterminals.*

Proof:

Let $\Sigma = [n]$ and $L_n = PERM(\Sigma)$.

1) Let $G = (N, \Sigma, S, P)$ be a Chomsky Normal Form Grammar for L_n. We assume that every element of N is used in some derivation of an element of L_n. We show that $|N| = \Omega\left(\frac{1.89^n}{n^{3/2}}\right)$.

Def 2.5 If A is a nonterminal then $GEN(A) = \{w \mid A \Rightarrow w\}$.

Claim 1: For all nonterminals A there exists a set $F \subseteq [n]$ such that $GEN(A) \subseteq PERM(F)$.

Proof of Claim 1:

Let $v, v' \in GEN(A)$. Then there exists u, x, u', x' such that

$$S \Rightarrow uAx \Rightarrow uvx \in PERM(\Sigma)$$

and

$$S \Rightarrow u'Ax' \Rightarrow u'v'x' \in PERM(\Sigma).$$

Clearly we also have

$$S \Rightarrow u'vx' \in PERM(\Sigma).$$

Hence v and v' must contain exactly the same letters (though they may be in a different order). Let F be the set of letters in v. Clearly $GEN(A) \subseteq PERM(F)$.

End of Proof of Claim 1

Def 2.6 If A is a nonterminal then let $F(A)$ be the set F proven to exist in the above claim.

Def 2.7 Let $w \in L_n$ and let T be a the parse tree for $w \in L(G)$. Let A be a nonterminal that apperas in the tree. Then $LE(A)$ is the set of leaves that are in the tree below A.

Claim 2: Let $w \in L_n$. There exists $(A, u, v, x) \in N \times \Sigma^* \times \Sigma^* \times \Sigma^*$ such that $w = uvx$, $v \in GEN(A)$, and $n/3 \leq |v| \leq 2n/3$.

Proof of Claim 2:

Look at the Parse tree for w. Since G is in Chomsky Normal Form the parse tree is binary. Start at the root. At every decision point goto the side that has the most leaves. Let B be the label on the first node such that the $LE(B) \leq n/3$. Let A be the parent of B. A has two children B and C. Note that $LE(A)$ has MORE THAN $n/3$ nodes below it since B is the FIRST node that has $LE(B) \leq n/3$ nodes below it. Also note that since $LE(B) \leq n/3$ and $LE(C) \leq LE(B)$,

$LE(C) \leq n/3$. Hence $LE(A) = LE(B) + LE(C) \leq 2n/3$. Hence Its easy to see that $n/3 \leq LE(A) \leq 2n/3$. Let v be the word generated by A in this parse Clearly $n/3 \leq |v| \leq 2n/3$.

End of Proof of Claim 2

Let N be the set of nonterminals of G. We map L_n to $N \times [n]$. Given $w \in L_n$ find (A, u, v, x) as in Claim 2. Let $i = |u| + 1$, so i is where the v-part starts. Map w to (A, i).

We upper bound the size of the inverse image of any $(A, i) \in N \times [n]$ and then use that to lower bound $|N|$.

Let $(A, i) \in N \times [n]$. How many w can map to it? Let $w = uvx$ where v begins at the ith spot and $n/3 \leq |v| \leq 2n/3$. Note that all of the w's that map to (A, i) have the same $|v|$, namely $|F(A)|$. We denote this by r and note that $n/3 \leq r \leq 2n/3$.

$v \in PERM(F(A)$. There are at most $r!$ such strings. The ux must be a perm of union of the letters in u and the letters in x. Hence $ux \in PERM(\Sigma - F(A))$. There are $(n - r)!$ such strings. Hence there are at most $r!(n - r)!$ elements mapping to (A, i). This is maximized when $r = n/3$ (or $r = 2n/3$). So each element of $N \times [n]$ has at most $(n/3)!(2n/3)!$ elements in the inverse image. Hence we get

$$n! \leq \sum_{A \in N, i \in [n]} (n/3)!(2n/3)! \leq |N|n(n/3)!(2n/3)!$$

Hence

$$|N| \geq \frac{1}{n} \frac{n!}{(n/3)!(2n/3)!}$$

By Lemma 2.2 $\frac{n!}{(n/3)!(2n/3)!} = \Theta(\frac{1}{\sqrt{n}} \frac{1}{(1/3)^{1/3}(2/3)^{2/3}}) = \Theta(\frac{1.89^n}{\sqrt{n}})$.

Hence

$$|N| \geq \Theta\left(\frac{1.89^n}{n^{3/2}}\right).$$

2) We give a CSL for L that has $O(n^2)$ nonterminals.

$$\begin{aligned}
S &\rightarrow A_1 A_2 \cdots A_n \\
A_i A_j &\rightarrow A_j A_i \text{ for all } 1 \leq i < j \leq n \\
A_1 &\rightarrow 1 \\
A_2 &\rightarrow 2 \\
&\vdots \\
A_n &\rightarrow n
\end{aligned}$$

This CSL is not in Chomsky Normal Form; however, it is easy to convert it to such without changing the number of nonterminals by too much. ▮

References

[EKSW05] K. Ellul, B. Krawetz, J. Shallit, and M. Wang. Regular expressions: new results and open problems. *Journal of Automata, Languages, and Combiatorics*, 10:407–437, 2005.