CS481F01 HW Solutions 5

A. Demers

26 Oct

1. Problem 73 from p. 334 of the text: Let $\Sigma = \{0, 1\}$. Let $\bar{x}$ denote the Boolean complement of x; that is, the string obtained from x by changing all $0's$ to $1's$ and all $1's$ to $0's$. Let x^r denote the reverse of x; that is, the string x written backwards. Consider the set

$$A \quad = \quad \{\ x \mid x^r \ = \ \bar{x}\ \}$$

For instance, the strings 011001 and 010101 are in A but 101101 is not.

(a) Give a CFG generating this set.

(b) Give grammars in Chomsky and Greibach normal form for $A - \{\epsilon\}$.

(answer a) Note these strings are all even length, since if the length were odd the middle symbol would have to be equal to its complement. Even-length strings in A can be constructed by generating complementary symbols "outside-in" using the following CFG:

$$S \quad \rightarrow \quad 0S1 \ \mid \ 1S0 \ \mid \ \varepsilon$$

An easy induction shows that

$$S \ \rightarrow^* \ \alpha S \beta \quad \Leftrightarrow \quad (\exists x)((\alpha = x) \ \wedge \ (\beta = (\bar{x})^r))$$

and correctness of the grammar follows.

(answer b) This language is simple enough that you can easily come up with CNF and GNF grammars directly, without applying the time-consuming conversion algorithm. A CNF grammar is just

$$\begin{aligned}
S &\rightarrow AC \mid DB \\
A &\rightarrow 0 \\
B &\rightarrow 1 \\
C &\rightarrow SB \\
D &\rightarrow AS
\end{aligned}$$

Since the nonterminals A, B, C, and D have only one production each, the equivalence of this grammar to the previous one (without ϵ) should be evident.

A GNF grammar is if anything easier:

$$\begin{aligned}
S &\rightarrow 0SB \mid 1SA \mid 0B \mid 1A \\
A &\rightarrow 0 \\
B &\rightarrow 1
\end{aligned}$$

Again, since A and B have only one production each, the equivalence to the original grammar should be clear.

2. Prove the following closure properties, where L and L' are CFLs and h is a homomorphism:

(a) (union): $L \cup L'$ is a CFL.

(b) (concatenation): $L \cdot L' = \{ xy \mid x \in L, y \in L' \}$ is a CFL.

(c) (homomorphic image): $h(L) = \{ h(x) \mid x \in L \}$ is a CFL.

(d) (inverse homomorphism): $h^{-1}(L) = \{ x \mid h(x) \in L \}$ is a CFL.

(answers a-c) These are fairly straightforward. Let

$$\begin{aligned}
G &= (N, \Sigma, P, S) & L(G) &= L \\
G' &= (N', \Sigma, P', S') & L(G') &= L'
\end{aligned}$$

For part (a), we construct

$$\begin{aligned}
G_a &= (N_a, \Sigma, P_a, S_a) \\
&\text{where} \\
N_a &= N \cup N' \cup \{S_a\} \\
P_a &= P \cup P' \cup \{S_a \to S\} \cup \{S_a \to S'\}
\end{aligned}$$

and for part (b) we construct

$$\begin{aligned}
G_b &= (N_a, \Sigma, P_a, S_a) \\
&\text{where} \\
N_b &= N \cup N' \cup \{S_b\} \\
P_b &= P \cup P' \cup \{S_b \to SS'\}
\end{aligned}$$

Finally, for part (c) we assume G is in GNF, and construct the grammar

$$\begin{aligned}
G_c &= (N, \Sigma, P_c, S) \\
&\text{where} \\
P_c &= \{A \to h(a)B_1/ldotsB_k \mid A \to aB_1 \ldots B_k \in P\ \}
\end{aligned}$$

Note the resulting grammar is not necessarily in GNF, and may even have ϵ-rules.

(answer d) This question is harder than I had remembered, and I apologize for it.

Note if the homomorphism h were a length-preserving homomorphism from Γ to Σ^* then this would be easy – the following simple construction would work.

$$\begin{aligned}
G_d &= (N, \Gamma, P_d, S) \\
&\text{where} \\
P_d &= \{A \to aB_1 \ldots B_k \mid Arightarrowh(a)B_1 \ldots B_k \in P\ \}
\end{aligned}$$

Unfortunately, when it is possible that $h(a) = \epsilon$ or $|h(a)| > 1$ things can get "out of alignment" and the above simple construction does not work. Consider the following GNF grammar and homomorphism:

$$\begin{aligned}
S &\to aBS \mid aBC \\
B &\to b \\
C &\to c
\end{aligned}$$

$$h(0) = a \qquad h(1) = ba \qquad h(2) = bc \qquad h(3) = \varepsilon$$

Now consider the derivation

$$S \to aBS \to abS \to abaBC \to ababC \to ababc$$

Here the entire string $ababc$ has a preimage 012 – that is, $h(012) = ababc$. However, the "inner" occurrence of S in this derivation generates the string abc, which has *no* preimage under h. For this derivation, we have to generate the b that precedes S along with the yield of S itself – $h(12) = babc$. So the preimages of the yield of a nonterminal *depend on the context of the nonterminal* in a derivation. In addition, instances of the symbol 3, which is annihilated by h, can occur anywhere in a preimage.

To get a construction that works, we can encode a finite amount of context in our nonterminals, much as we encoded states in our nonterminals in the proof that CFLs are closed under intersection with a regular set. Let m be the maximum length of a "right side" in the definition of h:

$$m \quad = \quad \max_{a \in \Gamma} |h(a)|$$

and let

$$\Sigma^{*m} \quad = \quad \{ \ x \in \Sigma^* \mid |x| \leq m \ \}$$

that is, the strings of length at most m. Now, we construct a grammar

$$G_d \quad = \quad (N_d, \Gamma, P_d, S_d)$$

The nonterminals of G_d will be nonterminals of G with left and right context from Σ^{*m}:

$$N_d \quad = \quad (\Sigma^{*m} \ \times \ N \ \times \Sigma^{*m}) \quad = \quad \{ \ [xAy] \mid x, z \in \Sigma^{*m}, A \in N \ \}$$

Intuitively, the nonterminal $[xAy]$ in G_d will correspond to an occurrence of A in a derivation

$$S \ \to_G^* \ w_1 A w_3 \ \to_G^* \ w_1 w_2 w_3$$

where the left context x is what is "left over" from an attempt to find a preimage of w_1:

$$(\exists u \in \Gamma^*)(h(u)x \ = \ w_1)$$

We're shooting for an inductive construction, so from A and the left context of A we'd like to be able to determine the left context of whatever follows A in the derivation. That's what the right context is for – the right context y of $[xAy]$ will be the left context of whatever follows, and it will be computed by

$$(\exists v \in \Gamma^*)((h(v)y \;=\; xw_2)$$

Note if both x and y were empty this would reduce to

$$h(v) \;=\; w_2$$

so v is a preimage of (the yield of A with a little error (x and y) at the ends), with the error recorded in the context parts of the nonterminal.

With that as motivation, we can give the rules of P_d: we add

$$\begin{array}{l} [xAy] \quad \rightarrow \quad [x_0B_1x_1][x_1B_2x_2]\ldots[x_{k-1}B_kx_k] \\ \qquad \text{if} \\ A \;\rightarrow\; a[B_1][B_2]\ldots[B_k] \;\in\; P \\ xa = x_0 \qquad \text{and} \qquad x_k = y \end{array}$$

Here we append a to the left context x if it "fits," i.e. if $|xa| \leq m$. We also add

$$\begin{array}{l} [xAy] \quad \rightarrow \quad b[x_0B_1x_1][x_1B_2x_2]\ldots[x_{k-1}B_kx_k] \\ \qquad \text{if} \\ A \;\rightarrow\; a[B_1][B_2]\ldots[B_k] \;\in\; P \\ h(b) \neq \varepsilon \qquad \text{and} \qquad xa = h(b)x_0 \qquad \text{and} \qquad x_k = y \end{array}$$

This is similar to the first rule, but we append a character (b) to the preimage. Now $h(b)$ generates a nonempty prefix of xa; the suffix of xa that remains is is now shorter than m and so can be passed as left context to B_1.

Finally, we add rules of the form

$$\begin{array}{l} [xAy] \quad \rightarrow \quad b[xAy] \;\mid\; [xAy]b \\ \qquad \text{if} \\ h(b) \;=\; \varepsilon \end{array}$$

to deal with symbols that are annihilated by h.

A correctness proof for this construction would proceed by proving the claim

$$[xAy] \quad \rightarrow^* \quad u \qquad \Rightarrow \qquad (\exists w)(A \;\rightarrow^*_G w \quad \wedge \quad xw \;=\; h(u)y)$$

and its converse. Having come up with this God-awful construction, the proof is actually a pretty routine induction. It just requires a little care to deal with the ε rules.

Applying this construction to the example from above (showing only the non-terminals that are actually used), we get

$$
\begin{array}{lcl}
[S] & \to & 0[Bb][bS] \mid 0[Bb][bC] \\
[bS] & \to & 1[Bb][bS] \mid 1[Bb][bC \\
[Bb] & \to & \varepsilon \\
[bC] & \to & 2 \\
\\
[S] & \to & 3[S] \mid [S]3 \\
[bS] & \to & 3[bS] \mid [bS]3 \\
[Bb] & \to & 3[Bb] \mid [Bb]3 \\
[bC] & \to & 3[bC] \mid [bC]3
\end{array}
$$

which generates the right stuff.

3. In this question you will show that the CFLs are not closed under intersection.

(a) Show that the set

$$\{\ 0^i 1^{i+1} \mid i > 0\ \}^*$$

is a CFL.

(answer a) This language is generated by

$$
\begin{array}{lcl}
S & \to & AS \mid \varepsilon \\
A & \to & 0A1 \mid 1
\end{array}
$$

The correctness should be clear.

(b) Show that the set

$$\{\ 0^{i^2} \mid i > 0\ \}$$

is *not* a CFL.

(answer b) This is similar to a pumping lemma example we did in lecture. Call this set of perfect squares L_S, and suppose it were regular. By the pumping lemma we know there must exist k such that any string in L_S of length greater than k can be "pumped". So choose n such that $(n^2) \geq k$ and $(2n+1) > k$.

$$0^{n^2} \;=\; uvwxy \qquad \text{where } \{\; uv^iwx^iy \mid i \geq 0 \;\} \;\subseteq\; L_S$$

Let $a = |vx|$. The pumping lemma says $a \leq k$. Consider the "pumped" string with $i = 2$:

$$z \;\equiv\; uv^2wx^2y \;\equiv\; 0^{n^2+a}$$

By our choice of n, we get

$$(n+1)^2 \;=\; (n^2+2n+1) \;>\; (n+k) \;>\; (n+a)$$

Thus, z is not in L_S, yielding our contradiction.

(c) Now show that the same set

$$\{\; 0^{i^2} \mid i > 0 \;\}$$

is a homomorphic image of the intersection of two CFLs.

(answer c) By part (b), it suffices to give two CFLs for which the lengths of strings in their intersections are the perfect squares. That's actually not too hard. Using the same technique used for part (a), we can show that the two languages

$$\{\; 0^i12^i \mid i > 0 \;\}^* \qquad \text{and} \qquad 0^*1\{2^i0^{i+1}1 \mid i > 0\}^*2^*$$

are both CFLs. The strings in the intersection of these two languages are of the form

$$01200122\ldots0^i12^i\ldots0^n12^n$$

The lengths of these strings are

$$\sum_{i=1}^{n}(2i+1) \;=\; n(n-1) \;+\; n \;=\; n^2$$

as required. Now the homomorphism

$$h(0) \;=\; h(1) \;=\; h(2) \;=\; 0$$

gives us the required result.