Grammars (part I)

Prof. Dan A. Simovici

UMB

Language classes can be defined using abstract models of computation, such as dfas, transition systems, etc.
Now we are concerned with an alternative approach that uses certain transformers of words called grammars. These systems can analyze or generate words and, therefore, can be used to recognize or generate languages.

Definition

A *grammar* is a 4-tuple $G = (A_N, A_T, S, P)$ such that:

- A_N is an alphabet whose members are the non-terminal symbols of G;
- A_T is the alphabet of terminal symbols;
- Alphabets A_N and A_T are disjoint;
- S is a symbol in A_N called the initial symbol;
- P is the set of productions of G (defined next).

The set $A = A_N \cup A_T$ is the alphabet of the grammar.

Definition

A production in a grammar $G = (A_N, A_T, S, P)$ is a pair of words (γ, γ') such that γ contains at least one non-terminal symbol.

If $G = (A_N, A_T, S, P)$ is a grammar and $\pi = (\gamma, \gamma') \in P$ we use the notation $\gamma \to \gamma'$.

Definition

If π is the production $\gamma \to \gamma'$, and $\alpha, \beta \in A^*$ such that $\alpha = \alpha_1\gamma\alpha_2$ and $\beta = \alpha_1\gamma'\alpha_2$, we say that *$\alpha$ generates β by applying the production $\gamma \to \gamma'$*.

This is denoted by $\alpha \underset{\pi}{\Rightarrow} \beta$.

If $\alpha \underset{\pi}{\Rightarrow} \beta$ for some production $\pi \in P$ of the grammar $G = (A_N, A_T, S, P)$ we write $\alpha \underset{G}{\Rightarrow} \beta$.

Definition

The sequence $d = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ is referred to as a *derivation of β from α in the grammar G*, where n is the *length of the derivation* if the following conditions are satisfied:

- $\alpha = \gamma_0$ and $\gamma_n = \beta$, and
- $\gamma_i \underset{G}{\Rightarrow} \gamma_{i+1}$ for $0 \leqslant i \leqslant n-1$.

An alternative notation for a derivation $d = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ is

$$\gamma_0 \underset{G}{\Rightarrow} \gamma_1 \underset{G}{\Rightarrow} \cdots \underset{G}{\Rightarrow} \gamma_n.$$

When the grammar is understood from the context, we may omit the subscript G.
Also, if there exists a derivation in G of the word β starting with α, we shall write $\alpha \overset{*}{\underset{G}{\Rightarrow}} \beta$.

The relation $\overset{*}{\underset{G}{\Rightarrow}}$ on the set of words $(A_N \cup A_T)^*$ is reflexive, that is,

$$\alpha \overset{*}{\underset{G}{\Rightarrow}} \alpha$$

for every word in $(A_N \cup A_T)^*$. This is interpreted as the existence of a derivation of length 0 of α from itself.

Example

The set of productions of the grammar

$$G = (\{S, X, Y\}, \{a, b, c\}, S, P),$$

consists of the productions listed below:

$$\begin{array}{llllll} \pi_0 & : & S \to abc, & \pi_1 & : & S \to aXbc, \\ \pi_2 & : & Xb \to bX, & \pi_3 & : & Xc \to Ybcc, \\ \pi_4 & : & bY \to Yb, & \pi_5 & : & aY \to aaX, \\ \pi_6 & : & aY \to aa & & & \end{array}$$

Example Cont'd

The following sequence is a derivation in G:

$$
\begin{aligned}
S &\underset{\pi_1}{\Rightarrow} aXbc \underset{\pi_2}{\Rightarrow} abXc \underset{\pi_3}{\Rightarrow} abYbcc \\
&\underset{\pi_4}{\Rightarrow} aYbbcc \underset{\pi_5}{\Rightarrow} aaXbbcc \underset{\pi_2}{\Rightarrow} aabXbcc \\
&\underset{\pi_2}{\Rightarrow} aabbXcc \underset{\pi_3}{\Rightarrow} aabbYbccc \underset{\pi_4}{\Rightarrow} aabYbbccc \\
&\underset{\pi_4}{\Rightarrow} aaYbbbccc \underset{\pi_6}{\Rightarrow} aaabbbccc.
\end{aligned}
$$

Definition

A derivation $\alpha_0 \underset{G}{\Rightarrow} \alpha_1 \underset{G}{\Rightarrow} \cdots \underset{G}{\Rightarrow} \alpha_n$ in a grammar $G = (A_N, A_T, S, P)$ is *complete* if $\alpha_n \in A_T^*$.

If $S \overset{*}{\underset{G}{\Rightarrow}} \alpha$, we refer to α as a *sentential form* of G.

Definition

The language generated by a grammar $G = (A_N, A_T, S, P)$ is the set of words

$$L(G) = \{x \in A_T^* \mid S \overset{*}{\underset{G}{\Rightarrow}} x\}.$$

Clearly, every word in $L(G)$ is a sentential form of G that contains no nonterminal symbols.

Types of Productions

Definition

Let A_N, A_T be two disjoint alphabets. A production $\alpha \to \beta$ is

1. a *context-free production* on A_N, A_T if α consists of one nonterminal symbol X and $\beta \in (A_N \cup A_T)^*$;
2. a *context-sensitive production* if $\alpha = \alpha' X \alpha''$ and $\beta = \alpha' \gamma \alpha''$, where $X \in A_N$, $\alpha', \alpha'', \gamma \in (A_N \cup A_T)^*$ and $\gamma \neq \lambda$.

Notational Recall

For a grammar $G = (A_N, A_T, S, P)$ we denote

- words from $(A_N \cup A_T)^*$ by $\alpha, \beta, \gamma, \ldots$;
- word from A_T^* by $x, y, z, u, \ldots$;
- derivations can be written as

$$\gamma_0 \underset{G}{\Rightarrow} \gamma_1 \underset{G}{\Rightarrow} \cdots \underset{G}{\Rightarrow} \gamma_n;$$

if we wish to specify the productions used, the same can be denoted as

$$\gamma_0 \underset{\pi_1}{\Rightarrow} \gamma_1 \underset{\pi_2}{\Rightarrow} \cdots \underset{\pi_n}{\Rightarrow} \gamma_n;$$

Example

Let $A_N = \{X, Y, Z\}$ and let $A_T = \{a, b\}$. The following pairs are context-free productions over A_N, A_T:

$$\begin{aligned} \pi_0 &: X \to abXYa \\ \pi_1 &: Y \to \lambda \\ \pi_2 &: Z \to bba \end{aligned}$$

The production $\pi_3 : aYXb \to abXZXb$ is context-sensitive; note that π_3 involves replacing Y by bXZ when Y is surrounded by a at left and by Xb at the right, that is, Y occurs in the context of a and Xb.

- Context-free productions of the form $X \to \lambda$ are called *null productions* or *erasure productions*. The effect of $X \to \lambda$ is to erase the symbol X.
- A grammar without erasure productions is said to be λ-*free*.

Let $G = (A_N, A_T, S, P)$ be a grammar.

- Every grammar is a grammar of type 0.
- G is of *type 1* (or, is *context-sensitive*) if all its productions are context-sensitive with the possible exception of a production $S \to \lambda$; if P contains $S \to \lambda$, then S does not occur in the right member of any production of G.
- G is of *type 2* (or, is *context-free*) if all its productions are context-free.
- G is of *type 3* (or, is *regular*) if every production has the form $X \to uY$ or $X \to u$, where $X, Y \in A_N$ and $u \in A_T^*$.

Definition

A grammar G is *length-increasing grammar* if all its productions are length-increasing with the possible exception of a production $S \to \lambda$; if P contains $S \to \lambda$, then S does not occur in the right member of any production of G.

It is clear that every grammar of type 3 is also of type 2, every grammar of type 1 is also of type 0 and every context-sensitive grammar is also length-increasing.

Definition

Let $\mathcal{G}$ be a class of grammars. A language L is a *$\mathcal{G}$-language* if there is a grammar G in $\mathcal{G}$ such that $L(G) = L$.

L is a context-free language if there exists a context-free grammar G such that $L = L(G)$.
Similarly, K is a length-increasing language if there is a length-increasing grammar G_1 such that $K = L(G_1)$, etc.
We denote by $\mathcal{L}_i$ the class of languages generated by grammars of type i for $0 \leq i \leq 3$.
Clearly, we have $\mathcal{L}_3 \subseteq \mathcal{L}_2 \subseteq \mathcal{L}_0$ and $\mathcal{L}_1 \subseteq \mathcal{L}_0$. Actually, as we shall see later, we also have the inclusion $\mathcal{L}_2 \subseteq \mathcal{L}_1$, so

$$\mathcal{L}_3 \subseteq \mathcal{L}_2 \subseteq \mathcal{L}_1 \subseteq \mathcal{L}_0.$$

- The corresponding classes $\mathcal{L}_i$ of languages constitute the Chomsky hierarchy.
- The inclusions between classes will be shown to be strict.
- It is clear that every language in $\mathcal{L}_1$ is length-increasing. Actually, we shall prove that $\mathcal{L}_1$ coincides with the class of length-increasing languages.

Example

The language generated by the context-free grammar

$$G = (\{S\}, \{a, b\}, S, \{S \to \lambda, S \to aSb\})$$

is $\{a^n b^n \mid n \in \mathbb{N}\}$.
We prove by induction on $n \geq 0$ that $a^n b^n \in L(G)$ for every $n \in \mathbb{N}$.
The case $n = 0$ follows from the existence of the production $\pi_0 : S \to \lambda$ in G.
Suppose now that $a^n b^n \in L(G)$, so $S \overset{*}{\underset{G}{\Rightarrow}} a^n b^n$. Using the production $S \to aSb$ we obtain the derivation

$$S \underset{G}{\Rightarrow} aSb \overset{*}{\underset{G}{\Rightarrow}} aa^n b^n b = a^{n+1} b^{n+1},$$

which shows that $a^{n+1} b^{n+1} \in L(G)$.

Example (cont'd)

Conversely, we prove by induction on the length $m \geq 1$ of the derivation $S \overset{*}{\underset{G}{\Rightarrow}} x$ that x has the form $x = a^n b^n$ for some $n \in \mathbb{N}$.

If $m = 1$, $S \overset{*}{\underset{G}{\Rightarrow}} x$ implies $x = \lambda$ since $S \to \lambda$ is the single production that erases S. Therefore, $x = a^n b^n$ for $n = 0$.
Suppose that the statement holds for derivations of length m and let $S \overset{*}{\underset{G}{\Rightarrow}} x$ be a derivation of length $m + 1$. If we write the first step of this derivation explicitly we have

$$S \underset{G}{\Rightarrow} aSb \overset{*}{\underset{G}{\Rightarrow}} x,$$

so $x = ayb$, where $S \overset{*}{\underset{G}{\Rightarrow}} y$ is a derivation of length m. By the inductive hypothesis, $y = a^n b^n$ for some $n \in \mathbb{N}$, so $x = a^{n+1} b^{n+1}$, which concludes our argument. Thus, $\{a^n b^n \mid n \in \mathbb{N}\}$ is a context-free language.

A Previous Example

Example

Consider again the length-increasing grammar

$$G = (\{S, X, Y\}, \{a, b, c\}, S, P),$$

where P consists of the following productions:

$$\begin{array}{llllll} \pi_0 & : & S \to abc, & \pi_1 & : & S \to aXbc, \\ \pi_2 & : & Xb \to bX, & \pi_3 & : & Xc \to Ybcc, \\ \pi_4 & : & bY \to Yb, & \pi_5 & : & aY \to aaX, \\ \pi_6 & : & aY \to aa & & & \end{array}$$

We claim that $L(G) = \{a^n b^n c^n \mid n \in \mathbb{P}\}$.

Example (cont'd)

- Any word $\alpha \in \{S, X, Y, a, b, c\}^*$ that occurs in a derivation, $S \overset{*}{\Rightarrow} \alpha$ contains at most one nonterminal symbol.
- A derivation must end either by applying the production $S \to abc$ or the production $aY \to aa$ because only these productions allow us to eliminate a nonterminal symbol.
- If the last production applied is $S \to abc$, then the derivation is $S \Rightarrow abc$, and the derived word has the form prescribed. Otherwise, the symbol Y must be generated starting from S, and the first production applied is $S \to aXbc$.

Example (cont'd)

For every $i \geq 1$ we have $a^i X b^i c^i \overset{*}{\Rightarrow} a^{i+1} X b^{i+1} c^{i+1}$. Indeed, we can write:

$$\begin{array}{lllll} a^i X b^i c^i & \underset{\pi_2}{\overset{i}{\Rightarrow}} & a^i b^i X c^i & \underset{\pi_3}{\overset{1}{\Rightarrow}} & a^i b^i Y b c^{i+1} \\ & \underset{\pi_4}{\overset{i}{\Rightarrow}} & a^i Y b^{i+1} c^{i+1} & \underset{\pi_5}{\overset{1}{\Rightarrow}} & a^{i+1} X b^{i+1} c^{i+1} \end{array}$$

We claim that a word α contains the infix aY (which allows us to apply the production π_5) and $S \overset{*}{\Rightarrow} \alpha$ if and only if α has the form $\alpha = a^i Y b^{i+1} c^{i+1}$ for some $i \geq 1$.

Example (cont'd)

An easy argument by induction on $i \geq 1$ allows us to show that if $\alpha = a^i Y b^{i+1} c^{i+1}$ then $S \overset{*}{\Rightarrow} \alpha$. We need to prove only the inverse implication. This can be done by strong induction on the length $n \geq 3$ of the derivation $S \overset{*}{\Rightarrow} \alpha$.

The shortest derivation that allows us to generate the word containing the infix aY is

$$S \Rightarrow aXbc \Rightarrow abXc \Rightarrow abYbcc \Rightarrow aYb^2c^2,$$

and this word has the prescribed form.

Example (cont'd)

Suppose now that for derivations shorter than n the condition is satisfied, and let $S \overset{*}{\underset{G}{\Rightarrow}} \alpha$ be a derivation of length n such that α contains the infix aY. By the inductive hypothesis the previous word in this derivation that contains the infix aY has the form $\alpha' = a^j Y b^{j+1} c^{j+1}$. To proceed from α' we must apply the production π_5 and replace Y by X. Thus, we have

$$S \overset{*}{\underset{G}{\Rightarrow}} a^j Y b^{j+1} c^{j+1} \underset{G}{\Rightarrow} a^{j+1} X b^{j+1} c^{j+1}.$$

Next, X must "travel" to the right using π_2, transform itself into an Y (when in touch with the cs) and Y must "travel" to the left to create the infix aY. This can happen only through the application of π_3 and π_4:

$$\begin{aligned} a^{j+1} X b^{j+1} c^{j+1} &\overset{j+1}{\underset{\pi_2}{\Rightarrow}} a^{j+1} b^{j+1} X c^{j+1} \\ &\overset{1}{\underset{\pi_3}{\Rightarrow}} a^{j+1} b^{j+1} Y b c^{j+2} \overset{i}{\underset{\pi_4}{\Rightarrow}} a^{j+1} Y b^{j+2} c^{j+2}, \end{aligned}$$

so α has the desired form. Therefore, the words in the language $L(G)$ have the form $a^n b^n c^n$.

There are certainly an infinite number of grammars for any language $L \in \mathcal{L}_0$.

Definition

Two grammars G, G' are *equivalent* if $L(G) = L(G')$.

There are many benefits to examining equivalent grammars that generate a language.

- For example, we may be given a context-sensitive grammar for a language for which there exists a context-free grammar or even a regular grammar. The simpler grammar leads to easier recognition of words in the language and provides more information about the structure of the language.
- By selecting specific characteristics of the form of the productions of a grammar, we may prove interesting facts about the language it generates.

Theorem

Let $G = (A_N, A_T, S, P)$ be a length-increasing grammar or a grammar of type i, where $i \in \{0, 1, 2\}$. There exists an equivalent grammar $G' = (A'_N, A_T, S, P')$ of the same type as G such that every production of P' that contains a terminal symbol is of the form $X \to a$.

Proof

Consider the alphabet $A' = \{X_a \mid a \in A_T\}$ that contains a symbol X_a for every terminal symbol a, where $A_N \cap A' = \emptyset$, and define A'_N as $A'_N = A_N \cup A'$.
The productions of P' are obtained by replacing each terminal symbol a by the corresponding nonterminal X_a and by adding the productions $X_a \to a$ for $a \in A_T$. The set of productions P' satisfies the requirements of the theorem, and the resulting grammar is clearly of the same type as G.

Proof (cont'd)

Let $u = a_{i_0} \cdots a_{i_{n-1}} \in L(G)$. The definition of the grammar G' implies that $S \overset{*}{\underset{G'}{\Rightarrow}} X_{a_{i_0}} \cdots X_{a_{i_{n-1}}}$. By using the productions $X_a \to a$ we obtain

$$S \overset{*}{\underset{G'}{\Rightarrow}} a_{i_0} \cdots a_{i_{n-1}},$$

so $a_{i_0} \cdots a_{i_{n-1}} \in L(G')$. Thus, $L(G) \subseteq L(G')$.

Proof (cont'd)

To prove the converse inclusion, $L(G') \subseteq L(G)$, consider a morphism $h : (A'_N \cup A_T)^* \longrightarrow (A_N \cup A_T)^*$ defined by $h(X_a) = a$ for $a \in A_T$ and $h(Y) = Y$ for every $Y \in A_N \cup A_T$. We claim that if $\alpha \underset{G'}{\Rightarrow} \beta$ for some $\alpha, \beta \in (A'_N \cup A_T)^*$, then $h(\alpha) \overset{*}{\underset{G}{\Rightarrow}} h(\beta)$. Indeed, if a production of the form $X \to a$ was used in $\alpha \underset{G'}{\Rightarrow} \beta$, then $h(\alpha) = h(\beta)$.

Proof (cont'd)

If another kind of production was used, then $h(\alpha) \underset{G}{\Rightarrow} h(\beta)$, so in any case, $h(\alpha) \overset{*}{\underset{G'}{\Rightarrow}} h(\beta)$. Let now $v \in L(G')$. We have $S \overset{*}{\underset{G'}{\Rightarrow}} v$, so $S = h(S) \overset{*}{\underset{G}{\Rightarrow}} h(v) = v$, which implies $v \in L(G)$. Therefore, $L(G) = L(G')$.

Example

Let $G = (\{S, X, Y\}, \{a, b, c\}, S, P)$ be the length-increasing grammar that generates the language $\{a^n b^n c^n \mid n \geqslant 1\}$. The grammar G' defined below is length-increasing, equivalent to G, and every production of this grammar that contains a terminal symbol is of the form $X \to a$. Specifically, the set of productions P' of the grammar

$$G' = (\{S, X, Y, X_a, X_b, X_c\}, \{a, b, c\}, S, P')$$

consists of the following productions:

$$\begin{array}{llllll}
\pi'_0 & : & S \to X_a X_b X_c, & \pi'_1 & : & S \to X_a X X_b X_c, \\
\pi'_2 & : & X X_b \to X_b X, & \pi'_3 & : & X X_c \to Y X_b X_c X_c, \\
\pi'_4 & : & X_b Y \to Y X_b, & \pi'_5 & : & X_a Y \to X_a X_a X, \\
\pi'_6 & : & X_a Y \to X_a X_a & \pi'_7 & : & X_a \to a \\
\pi'_8 & : & X_b \to b & \pi'_9 & : & X_c \to c
\end{array}$$