

BSc (Hons) Computer Science

University of Portsmouth

Third Year

Theoretical Computer Science

M21276

Semester 1

Hugh Baldwin

hugh.baldwin@myport.ac.uk

Contents

1	Lecture - Induction Lecture	2
2	Lecture - A1: Introduction to Languages	3
3	Lecture - A2: Grammars	7
4	Lecture - A3: Regular Languages	13
5	Lecture - A4: Finite Automata	17
6	Lecture - A5: Finite Automata & Regular Languages	23
7	Lecture - A6: Beyond Regular Languages	27
8	Lecture - A7: Pushdown Automata	29
9	Lecture - A8: Applications of Context-Free Grammars	33

Lecture - Induction Lecture

13:00

25/09/24

Janka Chlebkova

Contact Time

- Lectures
 - Delivered by Janka
 - Slides and videos from Covid times available weekly on Moodle
- Tutorials
 - Delivered by either Janka or Dr Paolo Serafino
 - Worksheets available with solutions weekly on Moodle
 - 5 Groups
 - Work through the worksheet solutions
 - Kahoot quiz for revision
 - ‘Tutorial 0’ available as revision of concepts from DMAFP needed for this module

Assessments

- 50%, 90 minute In-Class Test covering Part A - November 20th, 2pm
- 50%, 90 minute Exam covering Part B (Only) - TB1 Assessment Period (January)
- Deferred assessment covers both Part A & B, so will be harder than the two separate exams

There are two past papers (paper-based, but still relevant to the computer-based test) available for each of the assessments. As with DMAFP, the solutions are supplied in the form of a crowd-sourced document that Janka will check over before the exam.

Resources

Lecture slides, videos and tutorial papers are released every week. All of the books on the reading list are available online through the links on Moodle.

Lecture - A1: Introduction to Languages

13:00

03/10/24

Janka Chlebikova

Languages

In this context, a **language** is a set of symbols which can be combined to create a list of acceptable **strings**. There are then also rules which tell us how to combine these strings together, known as **grammars**. The combination of an alphabet, list of valid strings and grammars is known as a language. There will be a formal definition later on.

Definition

An **Alphabet** is a finite, non-empty set of symbols.

For example, in the English language we would define the alphabet, A , as $A = \{a, b, c, d, \dots, x, y, z\}$. These symbols can then be combined to create a **string**.

Definition

A **String** is a finite sequence of symbols from the alphabet of a language.

With the alphabet A , we could have strings such as 'cat', 'dog', 'antidisestablishmentarianism', etc over the alphabet. A string with no symbols, and therefore a length of zero, is known as the **empty string**, and is denoted by Λ (capital lambda).

Definition

A **Language** over an alphabet (e.g. English over A) is a set of strings – including Λ – made up of symbols from A which are considered 'valid'. They could be valid as per a set of rules, or could be arbitrary as in most spoken languages.

If we have an alphabet, Σ , then Σ^* denotes the infinite set of all strings made up of symbols in Σ – including Λ . Therefore, a language over Σ is any subset of Σ^* .

Example

If $\Sigma = \{a, b\}$, then $\Sigma^* = \{\Lambda, a, b, aa, ab, ba, bb, \dots\}$

There are many languages which could be defined over this alphabet, but a few simple one are

- \emptyset (The empty set)
- $\{\Lambda\}$ (The set containing an empty string)
- $\{a\}$ (The set containing a)
- The infinite set $\Sigma^* = \{\Lambda, a, aa, aaa, \dots\}$

Combining Languages

There are three common ways of combining two languages to create a new language – union, intersection and product.

Union and Intersection

Since languages are sets of strings, they can be combined as you usually would for any other set with the usual operations, union and intersection.

Example

If $L = \{aa, bb, cc\}$ and $M = \{cc, dd, ee\}$, then
 $L \cap M = \{cc\}$, and
 $L \cup M = \{aa, bb, cc, dd, ee\}$

Product

To combine two languages L and M , we form the set of all **concatenations** of strings in L with strings in M . This is known as the **product** of the two languages.

Definition

To **Concatenate** two strings is to juxtapose them such that a new string is made by appending the second string to the end of the first.

Example

If we concatenate the strings ab and ba , the result would be $abba$.
 This can be represented using the function cat , such as $\text{cat}(ab, ba) = abba$

With this definition of concatenation, we can say that the product of two languages, L and M would be $L \cdot M$, where $L \cdot M = \{\text{cat}(s, t) : s \in L \text{ and } t \in M\}$

Example

If $L = \{a, b, c\}$ and $M = \{c, d, e\}$, then the product of the two languages, $L \cdot M$, is the language
 $L \cdot M = \{ac, ad, ae, bc, bd, be, cc, cd, ce\}$

Further on Products

It is simple to see that for any language, L , the following simple properties are true:

$$L \cdot \{\Lambda\} = \{\Lambda\} \cdot L = L$$

$$L \cdot \emptyset = \emptyset \cdot L = \emptyset$$

Commutativity

Aside from the above properties, the product of any two languages is **not** commutative, and therefore

$$L \cdot M \neq M \cdot L$$

Example

If $L = \{a, b\}$ and $M = \{b, c\}$, then the product $L \cdot M$ is the language

$$L \cdot M = \{ab, ac, bb, bc\}$$

but the product $M \cdot L$ is the language

$$M \cdot L = \{ba, bb, ca, cb\}$$

which are clearly not the same language, as the only common string is bb

Associativity

However, the product of any two languages is associative. This means that if we had any three languages, L , M and N , then

$$L \cdot (M \cdot N) = (L \cdot M) \cdot N$$

Example

If we have the three languages $L = \{a, b\}$, $M = \{b, c\}$ and $N = \{c, d\}$, then

$$\begin{aligned} L \cdot (M \cdot N) &= L \cdot \{bc, bd, cc, cd\} \\ &= \{abc, abd, acc, acd, bbc, bbd, bcc, bcd\} \end{aligned}$$

which is the same as

$$\begin{aligned} (L \cdot M) \cdot N &= \{ab, ac, bb, bc\} \cdot N \\ &= \{abc, abd, acc, acd, bbc, bbd, bcc, bcd\} \end{aligned}$$

Powers of Languages

If we have the language L , then the product $L \cdot L$ of the language is denoted by L^2 . The product L^n for every $n \in \mathbb{N}$ is defined as

$$\begin{aligned} L^0 &= \{\Lambda\} \\ L^n &= L \cdot L^{n-1}, \text{ if } n > 0 \end{aligned}$$

Example

If $L = \{a, b\}$ then the first few powers of L are

$$\begin{aligned} L^0 &= \{\Lambda\} \\ L^1 &= L = \{a, b\} \\ L^2 &= L \cdot L = \{aa, ab, ba, bb\} \\ L^3 &= L \cdot L^2 = \{aaa, aab, aba, abb, baa, bab, bba, bbb\} \end{aligned}$$

The Closure of a Language

If L is a language over Σ , then the **closure** of L is the language denoted by L^* , and the **positive closure** is language denoted by L^+ .

Definition

The **Closure** of the language L , L^* is defined as

$$L^* = L^0 \cup L^1 \cup L^2 \cup L^3 \cup \dots$$

and so, if $L = \{a\}$ then

$$\begin{aligned} L^* &= \{\Lambda\} \cup \{a\} \cup \{aa\} \cup \{aaa\} \cup \dots \\ &= \{\Lambda, a, aa, aaa, \dots\} \end{aligned}$$

Definition

The **Positive Closure** the language L , L^+ is defined as

$$L^+ = L^1 \cup L^2 \cup L^3 \cup L^4 \cup \dots$$

and so, if $L = \{a\}$ then

$$\begin{aligned} L^+ &= \{a\} \cup \{aa\} \cup \{aaa\} \cup \{aaaa\} \cup \dots \\ &= \{a, aa, aaa, aaaa, \dots\} \end{aligned}$$

It then follows that $L^* = L^+ \cup \{\Lambda\}$, but it's not necessarily true that $L^+ = L^* - \{\Lambda\}$.

Example

If our alphabet is $\Sigma = \{a\}$ and our language is $L = \{\Lambda, a\}$, then

$$L^+ = L^*$$

Properties of Closures

Based upon these definitions, you can derive some interesting properties of closures.

Example

If L and M are languages over the alphabet Σ , then

$$\begin{aligned} \{\Lambda\}^* &= \emptyset^* = \{\Lambda\} \\ L^* &= L^* \cdot L^* = (L^*)^* \\ \Lambda &\in L \text{ if and only if } L^+ = L^* \\ (L^* \cdot M^*)^* &= (L^* \cup M^*)^* = (L \cup M)^* \\ L \cdot (M \cdot L)^* &= (L \cdot M)^* \cdot L \end{aligned}$$

These will be explored more during the tutorial session for this week

The Closure of an Alphabet

Going back to the definition of Σ^* of the alphabet Σ , it lines up perfectly with the definition of a closure such that Σ^* is the set of all strings over Σ . This means that there is a nice way to represent Σ^* as follows

$$\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \dots$$

From this, we can also see that Σ^k denotes the set of all strings over Σ whose length is k .

Lecture - A2: Grammars

13:00

03/10/24

Janka Chlebikova

As discussed previously, you can define a language over an alphabet as a set of arbitrary strings that are considered 'valid', but you can also define a language using a **grammar**.

Definition

A **Grammar** is a set of rules used to define a language over an alphabet, by specifying the structure of valid strings in said language.

To describe the grammar of a language, two sets of symbols (alphabets) are required, terminals and non-terminals

Definition

Terminal symbols are those from which the actual strings which make up a language are derived, like the symbols discussed previously for a manually specified language. In the case of a spoken language, these would usually be lowercase letters, such as the latin alphabet used in English.

Definition

Non-Terminal symbols are 'temporary' symbols (fully disjoint from the set of terminal symbols) used to define the grammar's replacement rules. These must all be replaced by terminals before a production can be considered a valid string in the language. These are usually represented by uppercase letters, or letters from an alphabet other than that of the language.

Productions

Further to this, a grammar for the language L over Σ would typically consist of a set of productions (grammar rules) which allow you to produce the set of strings which make up L .

Definition

A **Production** is a rule which allows you to produce a string for a language. They are typically in the form

$$\alpha \rightarrow \beta$$

where α is a string of symbols from the set of terminals (Σ) and β is a string of symbols from the set of non-terminals.

You can read these rules in several ways – $\alpha \rightarrow \beta$ could be read as;

- replace α by β
- α produces β
- α rewrites as β
- α reduces to β

A Formal Definition

Formally, a grammar is defined by the following properties

1. An alphabet T of terminal symbols (This is the same as the alphabet of the language defined by this grammar)
2. An alphabet N of non-terminal symbols
3. A specific non-terminal symbol known as the **start symbol**, which is often S
4. A finite set of productions in the form $\alpha \rightarrow \beta$ where α and β are strings over the alphabet $N \cup T$

Every grammar must have the special non-terminal symbol known as a start symbol, and it must also have at least one production in which the left side consists of only the start symbol.

Example

Let G be a grammar defined by

- The set of terminals $T = \{a, b\}$
- A single non-terminal being the start symbol S
- The set of productions

$$S \rightarrow \Lambda, S \rightarrow aSb$$

or for shorthand,

$$S \rightarrow \Lambda \mid aSb$$

To get the full set of strings which are valid for this grammar, we need to perform a **derivation**

Derivations

Starting from any production where the left side consists of only the start symbol, we can go through a step by step process to generate all strings belonging to the language described by a grammar.

Example

Using the grammar G from the previous example, we could start with either of the two productions, since they both have only the start symbol S on the left-hand side. This means that the first step of our process would give us Λ and aSb .

In this example, the string aSb is a **sentential form** of the terminals $\{a, b\}$ and non-terminal S . From this point, to extend or generate any other strings, we need to perform a derivation.

Definition

If x and y are sentential forms, and $\alpha \rightarrow \beta$ is a production, then to replace α by β in $x\alpha y$ is to **Derive** a new string. This is denoted by writing

$$x\alpha y \Rightarrow x\beta y$$

Example

Going back to the grammar G again, since it contains the production $S \rightarrow aSb$, we could take our first step to generate the string aSb , then we can go another step further and derive $aaSbb$ from aSb , which means

$$aSb \Rightarrow aaSbb$$

Since we can use this production again and again, we could derive that

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow \dots$$

To represent this multi-step process, we have the following symbols to show how the derivation was constructed

- \Rightarrow derives in a single step
- \Rightarrow^+ derives in one or more steps
- \Rightarrow^* derives in zero or more steps

Example

Using the grammar G , we can show multiple derivations using these new symbols

$$S \Rightarrow \Lambda, S \Rightarrow aSb, \therefore S \Rightarrow^* ab$$

$$S \Rightarrow^* aaabbb$$

$$S \Rightarrow^* aaaaaaSbbbbbb$$

What is $L(G)$?**Definition**

If G is a grammar with the start symbol S and the set of terminals T , then the language generated by G is the set

$$L(G) = \{s \mid s \in T^* \text{ and } S \Rightarrow^+ s\}$$

That is to say that it's the set of all strings containing only terminal symbols which can be derived from the start symbol with one or more steps.

Example

With the grammar G ,

$$L(G) = \{\Lambda, ab, aabb, aaabbb, aaaaabbbb, \dots\}$$

Infinite Languages

Within an infinite language, there is no limit on the length of strings, and therefore also no limit on the number of steps needed to derive a string. If the grammar of a language has n productions, then any derivation with $> n + 1$ steps must use at least one production twice. Therefore, if a language is infinite, then a production or sequence of productions must be used repeatedly to construct the derivations of the grammar.

Example

The infinite language $\{a^n b \mid n \geq 0\}$ can be described by a grammar with the production $S \rightarrow b \mid aS$. To derive the string $a^n b$, you would apply the production $S \rightarrow aS$ n times over and then finish the derivation with the production $S \rightarrow b$.

With the production $S \rightarrow aS$, we can also say that “If S derives w , then it also derives aw ”

Recursion and Indirect Recursion**Definition**

A production is **Recursive** if its entire left side is contained within the right side.

Example

The production $S \rightarrow aSb$ is recursive, since the sentential form it produces contains itself.

Definition

A production is **Indirectly Recursive** if it is possible to derive a sentential form in two or more steps which contains the entire left side of the production.

Example

If a grammar contains the rules $S \rightarrow b \mid aA$ and $A \rightarrow c \mid bS$, then both productions are indirectly recursive, since

$$\begin{aligned} S &\Rightarrow aA \Rightarrow abS \\ A &\Rightarrow bS \Rightarrow baA \end{aligned}$$

Recursive Grammars**Definition**

A grammar is **Recursive** if it contains either a recursive or indirectly recursive production. *In order for a grammar to be infinite, it must first be recursive.*

Constructing Grammars

We’ve seen previously how to derive a language from the definition of a grammar, but we also need to be able to construct a grammar to produce a language. It is often very difficult and sometimes impossible to write down a grammar for a given language, and there are often multiple grammars which form the exact same language.

Finite Languages

In the case of a finite language, it is always possible to find a grammar which produces the language. This is because it is possible to simply make a grammar which contains a production for every string of the language, where $S \rightarrow w$ for every string w in the language.

Example

The finite language $\{a, b, c, abc\}$ over the alphabet $\{a, b, c\}$ can be described by the grammar

$$S \rightarrow a|b|c|abc$$

Infinite Languages

However, in the case of an infinite language, there is no universal method that will always produce a valid grammar. This means it is much harder to find a valid grammar, and often requires **combining grammars**.

Example

A simple example of an infinite language would be $\{\Lambda, a, aa, aaa, \dots\}$ which we can also write in the form $\{a^n : n \in \mathbb{N}\}$. We can then see that one grammar which describes this language is as follows;

- The set of terminals, $T = \{a\}$
- The non-terminal start symbol, S
- The productions $S \rightarrow \Lambda, S \rightarrow aS$

Combining Grammars

Suppose we have two languages, L and M for which there are known grammars. There are a few simple rules we can use to create the grammars which describe the languages $L \cup M$, $L \cdot M$ and L^* .

To create this 'composite' grammar, we can describe L and M with disjoint sets of non-terminals. If we have the start symbols of L and M as A and B respectively, i.e. $L : A \rightarrow \dots$, $M : B \rightarrow \dots$, we can then combine these as productions in another grammar to create the language.

Unions

If we need to produce the union of two languages, $L \cup M$, we start with the two productions $S \rightarrow A \mid B$, followed by the productions which make up L and M , using A and B as their respective start symbols.

Example

If we wanted to write the grammar that produces the language

$$K = \{\Lambda, a, b, aa, bb, aaa, bbb, \dots, a^n, b^n\}$$

we can realise that K is the union of two different languages,

$$L = \{a^n \mid n \in \mathbb{N}\}$$

$$M = \{b^n \mid n \in \mathbb{N}\}$$

and so we can write a grammar for K as

- $S \rightarrow A \mid B$ (Union rule)
- $A \rightarrow \Lambda \mid aA$ (L s grammar)
- $B \rightarrow \Lambda \mid bB$ (M s grammar)

Products

Similarly, if we need to produce the the product of two languages, $L \cdot M$, we start with the production $S \rightarrow AB$, followed by the productions which make up L and M , using A and B as their respective start symbols.

Example

If we wanted to write the grammar that produces the language

$$K = \{\Lambda, a, b, aa, ab, aaa, bb, \dots\}$$

we can realise that K is the product of the two languages,

$$L = \{a^n \mid n \in \mathbb{N}\}$$

$$M = \{b^n \mid n \in \mathbb{N}\}$$

and so we can write a grammar for K as

- $S \rightarrow AB$ (Product rule)
- $A \rightarrow \Lambda \mid aA$ (L s grammar)
- $B \rightarrow \Lambda \mid bB$ (M s grammar)

Closures

And finally, if we need to produce the closure of a language, L^* , we start with the production $S \rightarrow AS \mid \Lambda$ followed by the productions which make up L , using A as its start symbol.

Example

If we wanted to write the grammar that produces the language L^* where

$$L = \{aa, bb\}$$

and so

$$L^* = \{\Lambda, aa, bb, aaaa, aabb, bbbb, bbaa, \dots\}$$

we can write the grammar for L^* as

- $S \rightarrow AS \mid \Lambda$ (Closure rule)
- $A \rightarrow aa \mid bb$ (L s grammar)

Equivalent Grammars

Any given language could have many different grammars which produce the exact same thing. This means that grammars are not unique. We can also use this to simplify grammars.

Example

Take the grammar from the previous example,

$$S \rightarrow AS \mid \Lambda$$

$$A \rightarrow aa \mid bb$$

If we replace A in S with the right side of A , aa , we get the production $S \rightarrow aaS$. We can then do the same with the other production to get $S \rightarrow bbS$. Therefore, we can write this grammar in a simplified form as

$$S \rightarrow aaS \mid bbS \mid \Lambda$$

Lecture - A3: Regular Languages

13:00

10/10/24

Janka Chlebikova

A regular language is a simple sort of language that meets a few basic requirements. They are often used for pattern matching and in lexical analysers.

Definition

There are several ways of describing a regular language

- Languages inductively formed by combining simple languages
- Languages described by a regular expression
- Languages produced by a grammar with a special and restricted form
- Languages that can be accepted by a finite automata

Inductively Formed Regular Languages

We start with a very simple language or set thereof and build more complex ones by combining them in particular ways–

- **Basis** - \emptyset , $\{\Lambda\}$ and $\{a\}$ are all regular languages for all $a \in \Sigma$
- **Induction** - If L and M are regular languages, then $L \cup M$, $L \cdot M$ and L^* are also regular languages

The basis of this definition gives us the following four regular languages over the alphabet $\Sigma = \{a, b\}$:

$$\emptyset, \{\Lambda\}, \{a\}, \{b\}$$

All regular languages over Σ can be constructed by combining these four languages in various ways by recursively applying the union, product and closure operations.

Example

Is the language $\{a, ab\}$ regular?

Yes, since it can be constructed from the four basic regular languages using a product and union operation–

$$\begin{aligned} \{a, ab\} &= \{a\} \cdot \{\Lambda, b\} \\ &= \{a\} \cdot (\{\Lambda\} \cup \{b\}) \end{aligned}$$

We can also see from this that it is possible to construct any finite language in this way, and therefore **all finite languages are regular**. There are also many infinite languages which are regular.

Regular Expressions

If you wanted to find an algorithm which can determine if a string belongs to a particular regular language, you would most likely use a **regular expression**

Definition

A **Regular Expression** is basically a short-hand way of showing how a regular language is constructed from a base set of regular languages.

For each regular expression E , there is a regular language $L(E)$

Like the regular languages discussed previously, regular expressions can be manipulated inductively to form new regular expressions.

- **Basis** - Λ , \emptyset and a are regular expressions for all $a \in \Sigma$
- **Induction** - If R and E are also regular expressions, then (R) , $R + E$, $R \cdot E$ and R^* are also regular expressions

Example

A few of the regular expressions over the alphabet $\Sigma = \{a, b\}$ are

$$\begin{aligned} &\Lambda, \emptyset, a, b \\ &\Lambda + b, b^*, a + (b \cdot a) \\ &(a + b) \cdot a, a \cdot b^*, a^* \cdot b^* \end{aligned}$$

So that we don't have to include lots of parentheses, the operations have the following hierarchy–

- $*$ (Highest)
- \cdot
- $+$ (Lowest)

Example

The regular expression

$$a + b \cdot a^*$$

is the same as

$$(a + (b \cdot (a^*)))$$

We can also juxtapose languages rather than using \cdot where there is only one possible interpretation

Example

The regular expression

$$a + b \cdot a^*$$

is the same as

$$a + ba^*$$

Operators

In regular expressions, there are two binary operations ($+$ and \cdot), and one unary operation ($*$). These are closely related to the union, product, and closure operations over corresponding languages.

Example

The regular expression

$$a + bc^*$$

is more-or-less shorthand for the regular language

$$\{a\} \cup (\{b\} \cdot \{c\}^*)$$

From Expressions to Languages

There is a very simple substitution method to find the language described by a regular expression.

Example

Find the language described by the regular expression $a + bc^*$ —

$$\begin{aligned} L(a + bc^*) &= L(a) \cup L(bc^*) \\ &= L(a) \cup (L(b) \cdot L(c^*)) \\ &= L(a) \cup (L(b) \cdot L(c)^*) \\ &= \{a\} \cup (\{b\} \cdot \{c\}^*) \\ &= \{a\} \cup (\{b\} \cdot \{\Lambda, c, c^2, c^3, \dots\}) \\ &= \{a\} \cup \{b, bc, bc^2, bc^3, \dots\} \\ &= \{a, b, bc, bc^2, bc^3, \dots\} \end{aligned}$$

This approach can also be used to prove that a given language is regular.

Many infinite languages can easily be proven to be regular, by finding a regular expression that describes it. Not all infinite languages are regular however.

Regular expressions are also not necessarily unique, as they may represent the same language as another regular expression, such as $a + b$ and $b + a$ being different regular expressions, but both represent the language $\{a, b\}$

Definition

Regular expressions are said to be **Equal** if their languages are the same. If R and E are regular expressions where $L(R) = L(E)$, then they are equal and can be written as $R = E$.

Properties of Regular Expressions

There are many general equalities for regular expressions, which hold for any regular expressions R , E , and F , and can be proven by using the basic properties of languages and sets.

- Additive Properties

$$\begin{aligned} R + E &= E + R \\ R + \emptyset &= \emptyset + R = R \\ R + R &= R \\ (R + E) + F &= R + (E + F) \end{aligned}$$

- Product Properties

$$\begin{aligned} R\emptyset &= \emptyset R = \emptyset \\ R\Lambda &= \Lambda R = R \\ (RE)F &= R(EF) \end{aligned}$$

- Distributive Properties

$$\begin{aligned} R(E + F) &= RE + RF \\ (R + E)F &= RF + EF \end{aligned}$$

- Closure Properties

$$\begin{aligned} \emptyset^* &= \Lambda^* = \Lambda \\ R^* &= R^*R^* = (R^*)^* = R + R^* \\ R^* &= \Lambda + RR^* = (\Lambda + R)R^* \\ RR^* &= R^*R \\ R(ER)^* &= (RE)^*R \\ (R + E)^* &= (R^*E^*)^* = (R^* + E^*)^* = R^*(ER^*)^* \end{aligned}$$

Regular Grammars

There are several ways to characterize grammars which describe regular languages.

Definition

A **Regular Grammar** is one where all of the productions take one of the following forms

$$\begin{aligned} B &\rightarrow \Lambda \\ B &\rightarrow w \\ B &\rightarrow A \\ B &\rightarrow wA \end{aligned}$$

Where A and B are non-terminals and w is a non-empty string of terminals

Only one non-terminal can appear on the right hand side of any production. Non-terminals must appear on the right end of the right hand side. Therefore the productions $A \rightarrow aBc$ and $S \rightarrow TU$ are not productions in a regular grammar, but $A \rightarrow abcA$ is.

For any regular language, we can find a regular grammar which describes it, but there may also be non-regular grammars which produce it.

Lecture - A4: Finite Automata

13:00

10/10/24

Janka Chlebikova

There are several different 'models of computation', such as finite automata, push-down automata and turing machines. They are all abstract models of computers which are capable of different things. They all have an input tape with a single string of an infinite length and can accept or reject the input string.

Finite Automata

Finite automata are the most basic models of a computer.

Definition

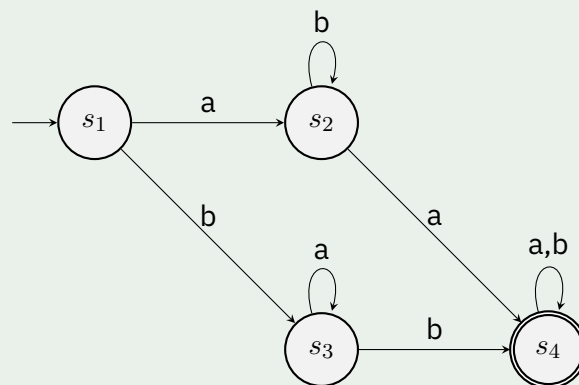
A **Finite Automaton** has three components–

- An input tape, which contains an input string over the alphabet Σ
- A head, which reads the input string one symbol at a time
- Memory, which is realised as a finite set Q of states, of which the automaton can only be in one at any time

The 'program' of the automaton defines how the read symbol changes the current state.

To better understand an automaton, they are typically represented as a transition graph, which is a form of directed graph.

Example



This is a finite automata which takes the alphabet $\Sigma = \{a, b\}$, and accepts strings such as *abba*, *baab*, etc.

Each automaton has one initial state, denoted by the arrow entering s_1 , and at least one final or accepting state, denoted by the double circle such as s_4 . Once the whole input string is read into the automaton, if the current state is a final or accepting state, then the string is accepted. Otherwise, the string is rejected. The language of an automaton is the set of strings it accepts.

A Formal Definition

Definition

A finite automaton is defined by

- A set of states, Q
- A start state, $s \in Q$
- A set of accepting states, $Q_a \subset Q$
- A set of transition functions $T : Q \times \Sigma \rightarrow Q$

Example

A very simple finite automaton could be defined by

$$Q = \{0, 1, 2\}$$

$$s = 0$$

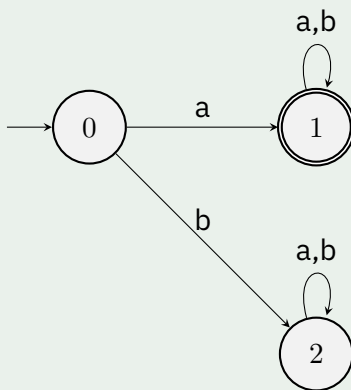
$$Q_a = \{1\}$$

$$T(0, a) = 1, T(0, b) = 2$$

$$T(1, a) = T(1, b) = 1$$

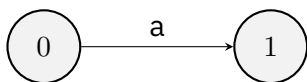
$$T(2, a) = T(2, b) = 2$$

And as a transition graph,



State Transition Functions

A state transition function, $T : Q \times \Sigma \rightarrow Q$, of the form



is represented by $T(0, a) = 1$, where $0, 1 \in Q$ and $a \in \Sigma$.

Deterministic Finite Automata

In the previous examples, there is always exactly one transition function for every state and symbol – every node has exactly one edge for each possible input symbol. Such automata are known as **deterministic** finite automata. That means that for any input string, we always know which unique state we are in at any symbol in the string.

Definition

A **DFA** accepts a string w over Σ^* if and only if there is a path from the initial state to an accepting state, such that w is the concatenation of the labels on the edges of the path. Otherwise, the DFA rejects w . For a FA to be deterministic, there must also be **exactly** one edge from each state for each symbol in Σ .

The set of all strings over Σ accepted by the DFA M is known as the language of M , and is represented by $L(M)$.

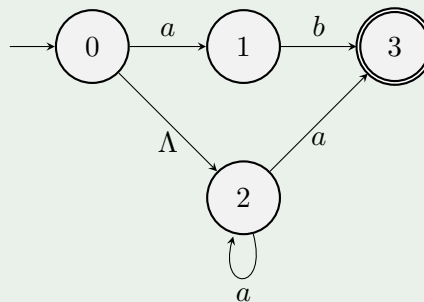
It has been proven that “the class of regular languages is exactly the same as the class of languages accepted by DFAs”. This means that for any given regular language, we can find at least one DFA which recognises it, and vice versa.

Non-Deterministic Finite Automata

With a DFA, you always know exactly which state it will be in after a given input string. This is not the case with an NFA, since in any given state there may be zero or more transitions for each symbol in the alphabet. Since more than one transition is possible for each state, it is impossible to know which state it will be in given an input string. If the NFA reaches an accepting state after reading the string, it is accepted. Otherwise it is rejected. Since there could be multiple paths for any given strings, as long as one or more path reaches an accepting state, it is accepted. Also, if there are no edges that can be traversed from a state using the current input symbol, the input string is immediately rejected.

Additionally, an edge may be labelled with Λ , which means that one can make the transition without consuming any symbols from the input string.

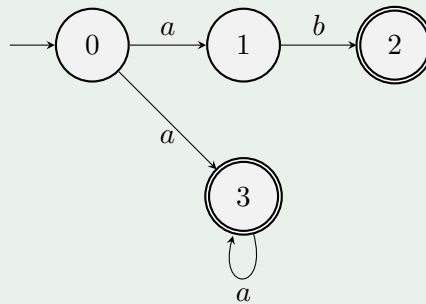
The non-determinism of these FA also means that they are impossible to run on a traditional computer, at least without using tricks that make them much less efficient. With the advent of quantum computers, it may become possible to use NFAs for much more efficient parsing and pattern matching.

Example

This is an NFA which corresponds to the regular expression $ab + a^*a$.

Since we are able to have multiple edges with the same symbol from a node, we could also use this different NFA which recognises the language of the same regular expression.

Example



NFA Transition Functions

Since NFAs are non-deterministic, their transition functions instead take the form $T : Q \times \Sigma \rightarrow P(Q)$. This allows for the result of any given transition function to be a set of states which could be transitioned to.

Example

Taking the first example of an NFA, the transition function might look as below–

$$\begin{aligned}
 T : Q \times \Sigma &\rightarrow P(Q) \\
 T(0, a) &= \{1\}, T(0, \Lambda) = \{2\} \\
 T(1, b) &= \{3\} \\
 T(2, a) &= \{2, 3\}
 \end{aligned}$$

and for the second example–

$$\begin{aligned}
 T : Q \times \Sigma &\rightarrow P(Q) \\
 T(0, a) &= \{1, 3\} \\
 T(1, b) &= \{2\} \\
 T(3, a) &= \{3\}
 \end{aligned}$$

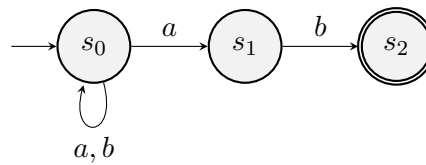
DFAs vs NFAs

On the surface it may seem like there aren't many differences between NFAs and DFAs, and in fact, DFAs are a subset of NFAs. It has also been proven that “the class of languages accepted by NFAs is also exactly the class of regular languages”, just as for DFAs. This also means that for every NFA there must be an equivalent DFA, or at least one that accepts the same language.

Generally, NFAs are easier to construct for any given regular expression, and tend to be more simple, with fewer states and transitions. However, DFAs are much easier to execute, especially on the deterministic computers we currently use. Since they recognise the same set of languages, it is always possible to find equivalents, though the DFA will usually be much larger and more complex.

Finding an Equivalent DFA

Generally, the DFA will have more states than the equivalent NFA. If the DFA has n states, the DFA may have as many as 2^n states. To actually do this conversion, you can follow the algorithm below (shown here converting the NFA in the figure below)

Figure 5.1: The NFA which accepts the language for the regular expression $(a + b)^*ab$

Step 1 – Begin with the NFA start state. If it's connected to any other states by Λ , make the initial state of the DFA a set containing the NFA's initial state and those others.

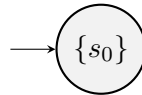


Figure 5.2: The resulting DFA after step 1

Step 2 – For each symbol, determine the set of possible NFA states you may be in after reading it. This set becomes the label of the state, and is connected to the previous state by that symbol. Only create a new state if one with that label does not already exist.

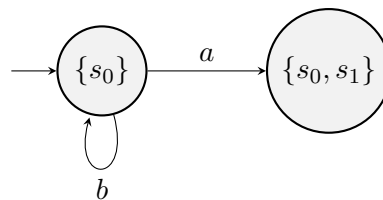


Figure 5.3: The resulting DFA after step 2

Step 3 – Repeat Step 2 for each new DFA state, until the system is closed. DFA accepting states are those that include one or more NFA accepting states. If there is no transition for a symbol in the NFA, create a new state in the DFA with the label \emptyset , and add loops for all symbols (known as a non-final trap state).

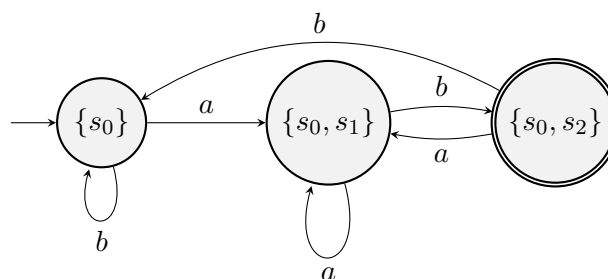


Figure 5.4: The final DFA

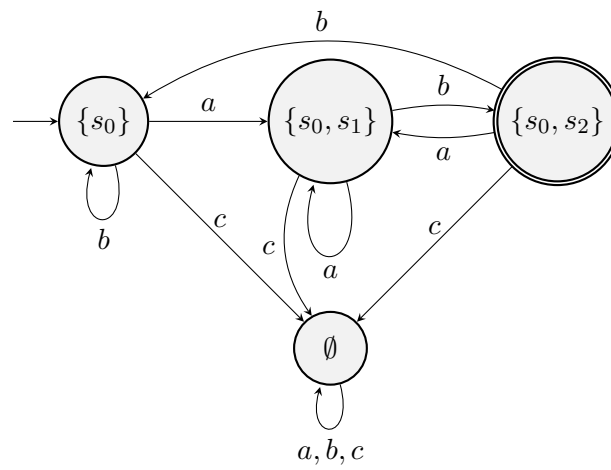


Figure 5.5: The same DFA, assuming that the language was $\{a, b, c\}$ rather than $\{a, b\}$

Lecture - A5: Finite Automata & Regular Languages

13:00

17/10/24

Janka Chlebkova

Automata and Regular Languages

How can we prove that the set of languages accepted by automata is exactly the set of regular languages? Well it takes multiple steps.

Step 1– We need to show that for any regular expression, we can find an NFA that recognises it, thus proving

$$L(\text{regular expressions}) \subseteq L(\text{NFA})$$

Step 2– Then we take a finite automaton and find a regular expression which describes the language of the automaton, thus proving that

$$L(\text{NFA}) \subseteq L(\text{regular expressions})$$

Step 3– We then combine the previous results to get that

$$L(\text{NFA}) = L(\text{regular expressions})$$

Regular Expressions to Finite Automata

Given any regular expression, we need to be able to construct a finite automaton (either NFA or DFA) which recognises its language. Given that all regular expressions are built up using union, product and closure operations, we can use a set of rules to construct an FA for a given regular expression.

Rule 0– Start the algorithm with a simple FA that has a start state, a single accepting state and an edge labelled with the given regular expression.

Rule 1– If an edge is labelled with \emptyset , erase the edge.

Rule 2– Transform any edge of the form $R + S$ into two edges, labelled R and S .

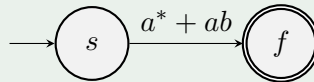
Rule 3– Transform any edge of the form $R \cdot S$ into a new intermediary state, with an edge labelled R going into it from the previous state, and an edge labelled S going into the next state.

Rule 4– Transform any edge of the form R^* into a new intermediary state with a looping edge labelled R , and two edges labelled Λ , one entering and one exiting the new state.

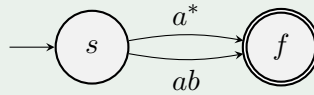
You must apply these rules until no more of the edges can be broken up any further using the rules.

Example

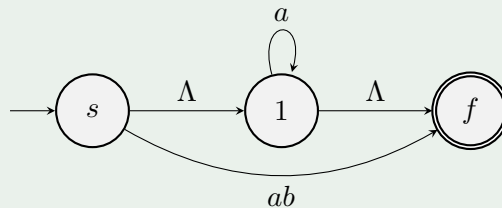
For the regular expression $a^* + ab$, we would follow the steps below—
Start with rule 0—



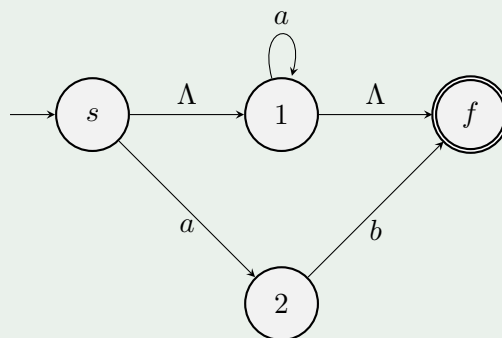
Apply rule 2—



Apply rule 4 to a^* —



Apply rule 3 to ab —



Thus, we have proven that we can produce a finite automata for any regular expression, and given a set of rules to do so.

Finite Automata to Regular Expressions

Now, to do the same thing in reverse, we need a new algorithm, as follows—

Step 1— Create a new start state s , and draw an edge labelled Λ to the original start state.

Step 2— Create a new accepting state f , and draw edges labelled Λ from all original accepting states.

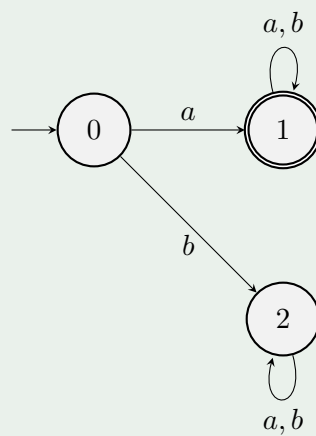
Step 3— For each pair of states i and j with more than one edge from i to j , replace all edges with a single edge labelled with the regular expression formed by the sum of labels on each edge from i to j .

Step 4— Step-by-step eliminate any states, changing the labels on corresponding edges until the only remaining states are s and f . When deleting a state, you must replace any possible transitions with a regular expression which matches all of the possible transitions.

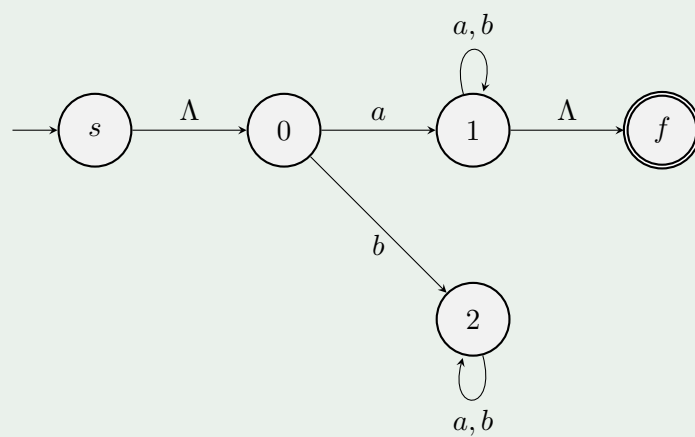
You should end up with an FA with only two states, and a single edge connecting them which is labelled with the desired regular expression.

Example

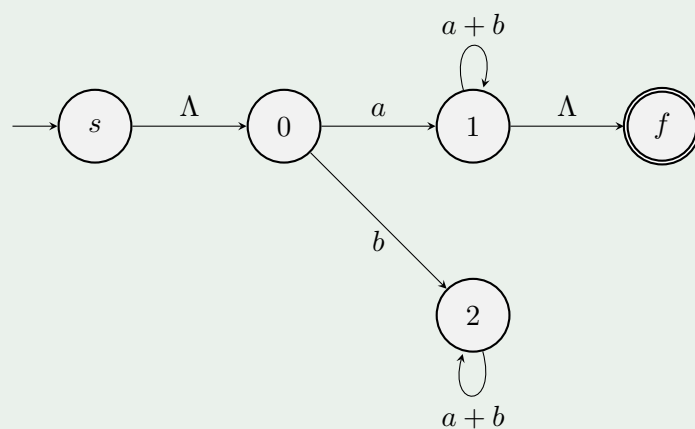
Initial DFA–



Steps 1 and 2: add new initial and final states–

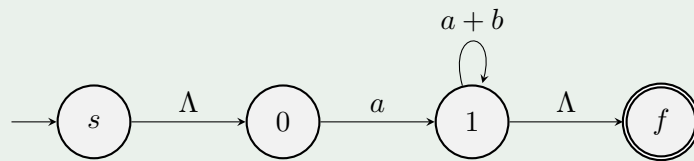


Apply step 3–

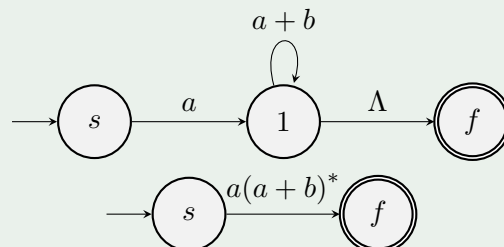


Example

We can then eliminate state 2, since there is no way for a string to be accepted from it–



Then we can start to eliminate states to get our final regular expression–



And this then leaves us with the final regular expression $a(a + b)^*$.

And now, we have proven that we can transform any regular expression into an NFA, and vice versa, and that we can transform any NFA into a DFA, and vice versa.

$$\text{regular expressions} \Leftrightarrow \text{NFA} \Leftrightarrow \text{DFA}$$

Simplifying DFAs

A lot of the time, we will end up with a complicated DFA, with far more states than is absolutely necessary. To get a simplified DFA, we need to transform it into a unique DFA with the minimum number of states which recognises the same language.

We can find this DFA in two parts– Finding all pairs of equivalent states, and combining the equivalent states into a single state, with modified transition functions.

Equivalent States

We can say that two states s and t are equivalent if for all possible strings w left to consume (including Λ), the DFA will finish in the same **type** of state (i.e. accepting or non-accepting) after consuming w . Therefore, once you arrive at s or t , the same string will always produce the same result (accepting or rejecting the string).

Two states are not equivalent if there exists a string w such that, after consuming w , we would end up accepting the string if we started from s and rejecting it if we started from t , or vice versa.

Lecture - A6: Beyond Regular Languages

13:00

17/10/24

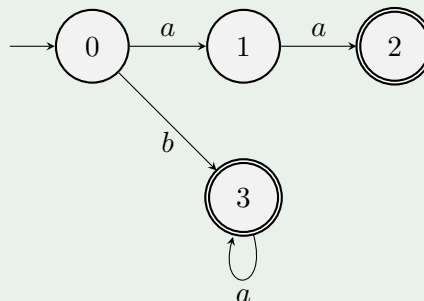
Janka Chlebková

Non-deterministic Finite Automata to Regular Grammars

Every NFA can be converted into a corresponding regular grammar. Each state of the NFA is associated with a non-terminal symbol of the grammar, and the initial state is associated with the start symbol. Each transition is associated with a grammar production, and every final state has the additional production which produces Λ .

Example

For the NFA–



–we can apply the rules defined above to get the following productions, supposing that $S = 0$, $A = 1$, $B = 2$ and $C = 3$

$$S \rightarrow aA \mid aC$$

$$A \rightarrow bB$$

$$B \rightarrow \Lambda$$

$$C \rightarrow aC \mid \Lambda$$

This may not be the simplest set of productions, but it does work.

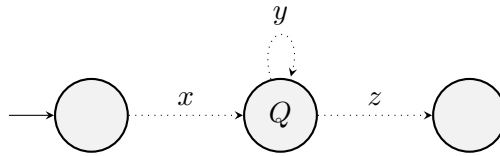
Testing for Regular Languages

Given a language, we need to be able to determine if it is regular. This would also allow us to prove that a given language cannot be recognised by any finite automaton.

One possible method is by using the **pumping lemma**, which applies for infinite languages. Since all finite languages are regular, this is enough to determine if a language is regular or not.

If the input string is long enough (e.g. greater than the number of states in the minimum state DFA for the language), then there must be at least one state Q which is visited more than once. Therefore, there must be at least one closed loop, which begins and ends at the state Q , and a particular string y , which corresponds to this loop.

We can represent this situation as the FA–



Where each dotted edge represents a path that may contain other states and transitions of the DFA.

- x is a string of symbols which the automaton consumes to transition from the start state, to get to the state Q
- y is the string of symbols to transition around the closed loop
- z is the string of symbols to transition from Q to an accepting state

Thus, we know that the string xyz is accepted, but also that the DFA must accept the strings $xy, xyz, xyxz, \dots, xy^kz$, and that the central string y is 'pumped'.

Formally, the pumping lemma is defined as follows

Definition

Let L be an infinite regular language accepted by a DFA with m states. Then any string w in L with at least m symbols can be decomposed as $w = xyz$ with $|xy| \leq m$ and $|y| \geq 1$ such that

$$w_i = xy^i z$$

is also in L for all $i = 0, 1, 2, 3, \dots$

This is necessary, but not sufficient for a language to be regular. If a language does not satisfy the pumping lemma, then it cannot be regular. But a language that satisfies the pumping lemma may not be regular. Therefore, you can use the lemma to prove that a language is **not** regular, but not that it is regular.

Lecture - A7: Pushdown Automata

13:00

24/10/24

Janka Chlebikova

There are many context-free languages that cannot be recognised by a finite automaton, especially non-regular languages. This is because they typically require an infinite memory to be recognised, such as the language $\{a^n b^n | n \geq 0\}$. A more powerful model of computing is known as Non-deterministic Pushdown Automata, or NDPA.

Non-Deterministic Pushdown Automata

An NDPA is like a finite automata, except that they also have memory in the form of a stack, where they can store an infinite amount of information. They are made up of the input tape, an infinite sequence of symbols from the input alphabet, a finite control automaton which is similar to an NFA, and the memory. Since the memory is modelled as a stack, it works in a Last In, First Out way, and there are 3 stack operations which can be performed at each step.

- pop– Reads the top symbol and removes it from the stack
- push– Writes a new symbol on to the top of the stack
- nop– No operation, has no effect on the stack

The symbols contained in the stack are different to the alphabet of the language the NDPA recognises. The initial state of the automaton is that the stack only contains the initial stack symbol, and that the control automaton is in its initial state.

Transition Functions

At each step, the current state, input symbol and symbol at the top of the stack determine the transition to the next state. Each transition must change the state, may read a symbol from and advance the input tape, and change the stack using one of the above stack operations.

Each transition function has three inputs

- The current state, p
- The input character, a or Λ
- The stack character, A

and two outputs

- The new state
- A stack operation

NDPA Formally

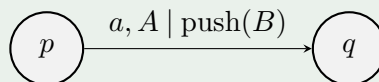
Definition

An NDPA can be formally described by

- A finite set Q of states, the initial state and the set of accepting states
- A finite set Σ of symbols making up the input alphabet
- A finite set Γ of symbols making up the stack alphabet, and the initial stack symbol, often $\$$
- A finite set of transition instructions, or a transition function T of the form $T : Q \times \Sigma \cup \{\Lambda\} \times \Gamma \rightarrow \Gamma^* \times Q$

Example

The transition



can also be written as the function

$$T(p, a, A) = (\text{push}(B), q)$$

or as the transition instruction

$$(p, a, A, \text{push}(B), q)$$

Instantaneous Descriptions of NDPAs

During computation, you need to be able to describe the NDPA after any given set of inputs. Since it also modifies the stack at each step, the entire stack needs to be included in the description. Formally, this is known as the **instantaneous description** of an NDPA.

Definition

The instantaneous description consists of

- The current state
- The unconsumed input characters
- The contents of the stack

and is written as

$$(\text{current state}, \text{unconsumed input}, \text{stack contents})$$

In this case, the top of the stack is on the left, and bottom on the right.

Example

$$(0, abba, YZ\$)$$

Accepted Languages

A string is accepted by an NDPA if there is some path from the initial state to an accepting state which consumes the entire input string. Otherwise, the string is rejected by the NDPA. The language is then the set of all strings that it accepts.

Specifically, an input string is rejected if it finishes reading the string before reaching an accepting state, if the current state has no transition for the current input or stack symbol, or finally if it attempts to pop an empty stack.

Example

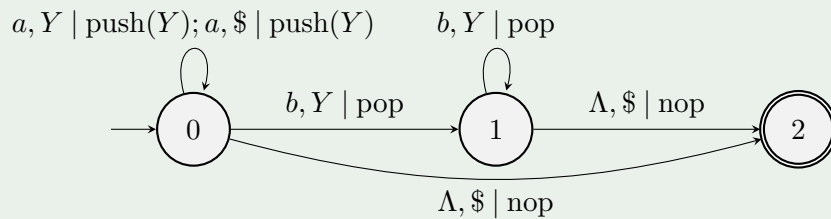
If we wanted to build an NDPA that recognises the language $\{a^n b^n | n \geq 0\}$, we could use the following plan.

- Read the string, and for each a that's read, push Y onto the stack
- Once the first b is read, change state and start removing one Y from the stack for each b read
- If the input string ends and the stack has just emptied, the string is accepted
- Otherwise, the string should be rejected

We can create this NDPA as the automaton with

- $Q = \{0, 1, 2\}$ where 0 is the initial state, and 2 accepting
- $\Sigma = \{a, b\}$
- $\Gamma = \{Y, \$\}$

and drawn as

**NDPAs and Context-Free Languages**

The class of all languages accepted by NDPAs is exactly the same as the class of context-free languages. To prove this, we need to show that

1. With any given NDPA, we can find a context-free grammar which produces the same language
2. With any given context-free language, we can find an NDPA which accepts the given language

Example

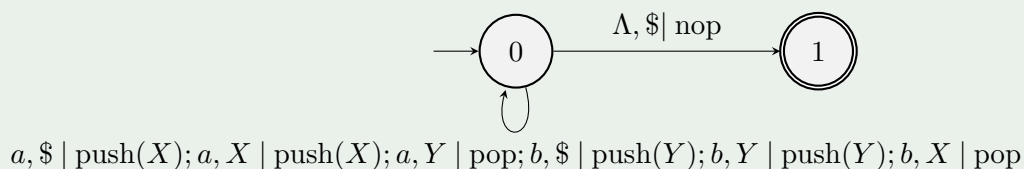
Show that the language containing all strings over $\{a, b\}$ with exactly the same number of a s and b s is context-free. Using the theorem, all we need to do is find an NDPA which recognises this language.

Since we need to keep track of the number of a s and b s, we can change the stack accordingly. If we've seen more a s, add an X , and if we've seen more b s, add a Y .

We can then represent this NDPA as

- $Q = \{0, 1\}$ where 0 is the initial state, and 1 accepting
- $\Sigma = \{a, b\}$
- $\Gamma = \{X, Y, \$\}$

with the transition diagram

**Determinism**

Similarly to finite automata, push-down automata can be either deterministic or non-deterministic. However, a non-deterministic push-down automata isn't very useful, since it can never make a choice at each state. Unlike NFAs and DFAs, deterministic push-down automata cannot recognise the entire family of

context-free languages.

Lecture - A8: Applications of Context-Free Grammars

13:00

24/10/24

Janka Chlebkova

The original purpose of context-free languages was to describe the grammar of English, in terms of its block structure, built up recursively from smaller phrases. However, since phrases with the exact same structure can have one of several meanings, we need a way to show which of the possible derivations was used to produce it. This can be shown using a parse (or derivation) tree, which starts with the initial symbol and works down until all of the leaves are terminal symbols (usually words in a spoken language).

- The start symbol of the grammar becomes the root of the tree
- For each production $X \rightarrow Y_1 \dots Y_n$, add $Y_1 \dots Y_n$ as children of the node X
- Every leaf must end up as a terminal, and any internal nodes must be non-terminals

If there are multiple possible parse trees for a given string in a language, how do we know which one is correct? Well, all of them are. This means that for any string with a non-unique parse tree, there are several correct meanings.

Ambiguous Grammars

If a string in a grammar has a non-unique parse tree, then it is known as **ambiguous**. This is because there would be no way for a compiler or other computer program to know which of the options is correct. As a human, you can usually figure it out based upon other context, but that is not a luxury afforded to computers.

Conversely, if each string of a language has only one parse tree (or alternatively, if there is only one left-most (or right-most) derivation for each string), then the language is unambiguous.

Since there are no general techniques for handling ambiguity, it is impossible to convert an ambiguous grammar into an unambiguous one. In some cases, a grammar can be modified such that it produces the same language, but is unambiguous, however this can only be done on a case-by-case basis. An example of this is adding parenthesis in mathematical expressions to make sure that the correct order of operations is used.

Parsing

Parsing is one of the main steps of a compiler, and is a process in which

- It is determined whether the input string has correct syntax
- A parse tree is built, which represents the internal structure of a string, and how it was derived from the grammar

Parsing can be done in one of two ways— top-down parsing and bottom-up parsing.

Top-down parsing constructs the parse tree by starting from the grammar's start symbol, and replacing non-terminals, working its way towards the string. If, after following every logical set of non-terminals, it is not possible to produce the input string, then the string cannot be parsed and it is not in the language of the parser.

Bottom-up parsing starts with the string, and works backward to get to the start symbol, replacing sets of terminals with a grammar rule which produces them. Once again, if all combinations fail then the string cannot be parsed and is not in the language of the parser.

When working with a top-down parser, there is often a need to backtrack through the parse tree, if it gets stuck at some point and cannot progress any further towards the final string, which means that the whole state at each step needs to be stored until the parse is completed. Some grammars are backtrack-free, and it is sometimes possible to design a grammar purposefully like that.

When bottom-up parsing, the internal state also needs to be stored, but there is usually less backtracking required, depending on the grammar. They can also handle a larger class of grammars than top-down parsers.

Computer Languages

Since a parser is one of the most important parts of a compiler, it is desirable to design the programming language in such a way that it is easy to construct a parser for it. This means that a lot of programming languages take the form of a context-free grammar, but more specifically there are categories of CFG that are easier for either a top-down or bottom-up parser to parse.

The best candidates for these easier languages are either $LL(k)$ - or $LR(k)$ -grammars, for any $k \geq 0$.

- $LL(k)$ -grammars are based on $LL(k)$ -parsers, and are better for top-down parsing
- $LR(k)$ -grammars are based on $LR(k)$ -parsers, and are better for bottom-up parsing

$LL(k)$ -grammars

What does $LL(k)$ actually mean?

- L – The input string is parsed from left-to-right
- L – Only left-most derivations are considered at each step
- k – The number of look-ahead symbols needed to decide which derivation to use

Example

The grammar $S \rightarrow aSc|b$ for the language $LL(1) = \{a^nbc^n, n \geq 0\}$ only requires one symbol of look-ahead, since it is always possible to know which production needs to be used to continue parsing any given string.

Example

The grammar $S \rightarrow AB, A \rightarrow aA|a, B \rightarrow bB|c$ requires two symbols of look-ahead, since there are multiple possible productions for any given symbol. If we had the input symbol a , we could have used either of the two productions of A .

There are some languages that cannot be represented by an $LL(k)$ grammar, such as if any of the productions start with a non-terminal, since you would need to know the length of the full string. That means that it is impossible to do a partial derivation, knowing only how the string starts.

All $LL(k)$ grammars are unambiguous, and are a subset of deterministic context-free languages. $LL(1)$ languages are very popular, as they are much easier to parse than other languages. Some languages however are too complex to be described by an $LL(1)$ grammar, such as the C language, however these can be parsed using higher order grammars.

$LR(k)$ -grammars

Languages based upon $LR(k)$ grammars are parsed bottom up, scanning the string from left-to-right, as with $LL(k)$ grammars, but produce a right-most derivation in reverse.

- L – The input string is parsed from left-to-right
- R – Only right-most derivations are considered at each step
- k – The number of look-ahead symbols needed to parse the grammar

Donald Knuth proved that all $LR(1)$ languages are exactly $LR(k)$ languages, which are exactly deterministic context-free languages (DCFL). DCFGs are a proper subset of the context-free grammars, and can be recognised by DPDAs. They are also always unambiguous.

They are of particular interest since they can be parsed in linear time, and a parser can be automatically generated from the grammar using a parser generator.

To derive an $LR(k)$ grammar, you start from the bottom up, essentially scanning from left-to-right, until you find the first right-hand side of a production. You then replace the symbols in the string with the left-hand side of the production, and continue the process on the new string. This continues until you eventually cannot find any more productions (in which case, the string does not belong to the language), or you reach the start symbol. You can then invert the order of derivations to find the original derivation of the string.