

UMEÅ UNIVERSITY
Efficient Algorithms

ASSIGNMENT STEP 2

Runtime analysis of the Java implementation of the CYK-algorithm

Pina Kolling

September 26, 2022

Contents

1	Introduction	2
2	Background	3
2.1	Formal language	3
2.2	Formal grammar	3
2.3	Chomsky-Normal-Form	3
2.4	CYK-algorithm	4
2.5	Dynamic programming	4
3	System Design	5
3.1	Main	5
3.2	Grammar	5
3.3	Parser	6
3.3.1	Naive	6
3.3.2	Bottom-Up	6
3.3.3	Top-Down	6
4	Evaluation	7
5	Conclusions and Future Work	8
A	How to use the code?	9
	Bibliography	10

1 Introduction

Parsing in Computer Science is the process of analysing a string of characters to examine if the string is built according to the rules of a formal grammar.

A formal grammar describes how to form strings with correct syntax from a language's alphabet (section 2.2 and 2.1). To examine if such a string follows the rules of a grammar the *Cocke-Younger-Kasami*-algorithm (short: *CYK*) can be used. This algorithm is described in section 2.4. To use the *CYK*-algorithm the grammar needs to be in a specific format, that is called *Chomsky-Normal-Form* (*CNF*), which is explained in section 2.3. [4, 6]

The task for this assignment was to code three different parsing methods to execute the *CYK*-algorithm in *Java*. The different parsing methods will be described and presented as pseudo code in section 3. For the implementation three different classes were implemented: `main.java`, `grammar.java` and `parser.java`. The `main`-class calls the methods and the `grammar`-class parses the input grammar and string into a format that then can be processed in the `parser`-class. The function and implementation will be further described in section 3.

2 Background

In this section background information on formal languages (section 2.1), formal grammars (section 2.2), Chomsky-Normal-Form (section 2.3), CYK-algorithm (section 2.4) and dynamic programming (section 2.5) will be presented.

2.1 Formal language

Formal languages are abstract languages which define the syntax of the words that get accepted by that language. It consists of a set of words that get accepted by the language and a set of symbols that is called alphabet and contains the characters of the words. Those characters are called nonterminal symbols. [1, 5]

Example (Formal language).¹

The language accepts words that contain the same number of a s and b s, while the a has to be left of the b . The alphabet Σ of this language looks like this:

$$\Sigma = \{a, b\}$$

The language definition L is the following one:

$$L = \{(a^n b^n)^m\} \text{ with } n, m \in \mathbb{N}$$

2.2 Formal grammar

A formal grammar describes how to form strings with correct syntax from a language's alphabet. A grammar does not describe the meaning of the strings or any semantics — only their syntax is defined. The grammar is a set of rules which define which words are accepted by a formal language. Those rules consist of terminal and nonterminal symbols. The terminalsymbols are the characters of the alphabet of the language and the nonterminalsymbols are used to build the rules of the language — they get replaced by terminalsymbols. [1, 5]

Example (Formal grammar). *The example grammar for the previous example language is the following:*

$$S \rightarrow SS \mid aSb \mid ab$$

2.3 Chomsky-Normal-Form

The *Chomsky-Normal-Form* (short: *CNF*) is a grammar which is formatted in a specific way. If the startsymbol (nonterminal symbol) is not generating the empty word ($S \rightarrow \epsilon$) it either generates two nonterminal symbols or one terminal symbol for the grammar to be in *CNF*. [1]

¹The following examples show the *Well-Balanced Parentheses* example from the assignment task sheet with the alphabet $\{a, b\}$ instead of $\{(\cdot, \cdot)\}$.

$$A \rightarrow SR$$

$$L \rightarrow a$$

$$R \rightarrow b$$

Cocke-Younger-Kasami-algorithm (short: *CYK*) takes a grammar in *CNF* and a word as an input. It then examines if the word follows the rules of the grammar. [6]

The word is accepted by the grammar rules, because the initiating nonterminal symbol S can be filled into the lowest field.

The technique of dynamic programming can be used to solve problems, which can be divided into smaller subproblems. The solutions of the subproblems are saved (for example in a multi-dimensional array) and referenced later. [2]

3 System Design

The implementation was done in Java and three different classes were implemented: `main.java` (described in section 3.1), `grammar.java` (described in section 3.2) and `parser.java` (described in section 3.3). The `main`-class calls the methods and the `grammar`-class parses the input grammar and string into a format that then can be processed in the `parser`-class.

3.1 Main

The `main`-class takes the input grammar and word and parses them into `String[]` and `String`. The arguments have to follow the following rules:

- The Grammar needs to be in *CNF*.
- The first rule begins with the start symbol of the grammar.
- The rules are put in without arrows, one rule body is represented by one string, beginning with the rule head.
- The last argument is the input word.

Input example (*Well-Balanced-Parantheses*):

```
java Main "SSS" "SLA" "SLR" "ASR" "L(" "R)" "((())"
```

for the grammar $S \rightarrow SS \mid LA \mid LR$, $A \rightarrow SR$, $L \rightarrow ($, $R \rightarrow)$ and the input word `((())`.

3.2 Grammar

The `grammar`-class assigns the nonterminal symbols to integers and builds arrays with them. The start symbol for example is then assigned with the integer zero and the bodies of that rule are at the index zero of a two-dimensional array. One array that only contains nonterminal symbols is built, one array that contains terminal and nonterminal symbol and one array that represents the integers that represent each nonterminal symbol is built.

For the input `java Main "SSS" "SLA" "SLR" "ASR" "L(" "R)" "((())"` the two-dimensional arrays shown on the screenshot on the right side are built.

(Right now the arrays still have the type `String[][]`, that will be changed later to `int[][]` to minimize the access time in the parsing methods.)

```
Matrix all rules:
[SS, LA, LR]
[SR, , ]
[(, , ]
[), , ]

Matrix T rules:
[]
[]
[()]
[]

Matrix NT rules:
[SS, LA, LR]
[SR, , ]

Integers of NT symbols:
[0, 1, 2, 3]
[S, A, L, R]

Matrix all rules:
[00, 21, 23]
[03, , ]
[(, , ]
[), , ]

Matrix T rules:
[]
[]
[()]
[]

Matrix NT rules:
[00, 21, 23]
[03, , ]
```

3.3 Parser

The `parser`-class contains three different parsing methods. Each methods has a counter as `long` which counts the number of iterations and a timer as `long` in *ms* which measures the runtime of the method.

3.3.1 Naive

The naive algorithm is a recursive algorithm and has an initial method to start the recursion with the start values. Then the recursive method gets called.

Algorithm 1 Recursion call: Boolean `parseNaive()`

```
1: counter  $\leftarrow$  0
2: return parseNaive(0, 0, inputWord.length)
```

Algorithm 2 Boolean `parseNaive(int indexNT, int i, int j)`

```
1: counter  $\leftarrow$  +1
2: if i == (j - 1) then
3:   for l  $\leftarrow$  0 to ruleset[0].length do
4:     if ruleset[indexNT][1] == inputWord[i] then
5:       return true
6:     end if
7:   end for
8: else
9:   for bodyIndex  $\leftarrow$  0 to ruleset[indexNT].length do
10:    if ruleset[indexNT][bodyIndex].length >= 2 then
11:      first  $\leftarrow$  ruleset[indexNT][bodyIndex].charAt(0)
12:      second  $\leftarrow$  ruleset[indexNT][bodyIndex].charAt(1)
13:      for k  $\leftarrow$  i + 1 to j do
14:        if parseNaive(first, i, k) and parseNaive(second, k, j) then
15:          return true
16:        end if
17:      end for
18:    end if
19:  end for
20: end if
21: return false
```

3.3.2 Bottom-Up

3.3.3 Top-Down

4 Evaluation

We did some experiments ...

5 Conclusions and Future Work

From our experiments we can conclude that ...

A How to use the code?

The code can be run in the terminal and input is expected as Strings in quotation marks. The grammar needs to be in CNF. The first rule begins with the startsymbol of the grammar.

First: Rules without arrows (one rule as one String)

Last: The last argument is the input word

Input example (*Well-Balanced-Parantheses*):

```
java Main "SSS" "SLA" "SLR" "ASR" "L(" "R)" "(())"
```

for the grammar $S \rightarrow SS \mid LA \mid LR$, $A \rightarrow SR$, $L \rightarrow ($, $R \rightarrow)$ and the input word $(())$.

Output example:

The first part of the output shows the arrays, which get generated in the `Grammar.java` class.

The first array contains all rules.

The second array contains only the terminal rules.

The third array contains only the nonterminal rules.

Then it is shown which nonterminal symbols are represented by which integers. Later the nonterminal symbols can be referred to with those integers.

After this the mentioned arrays are shown again but the nonterminal symbols got replaced with the according integers.

```
Input word: (()
Naive: true   Amount of calls: 33
Naive runtime: 9ms

CYK-Table (Bottom Up):
[2, , , ]
[ , 2, 0, 1]
[ , , 3, ]
[ , , , 3]

BottomUp: false   Amount of calls: 84
Naive runtime: 4ms

TopDown: true   Amount of calls: 28
Naive runtime: 1ms
```

```
Matrix all rules:
[SS, LA, LR]
[SR, , ]
[(, , ]
[), , ]

Matrix T rules:
[]
[]
[(]
[)]

Matrix NT rules:
[SS, LA, LR]
[SR, , ]

Integers of NT symbols:
[0, 1, 2, 3]
[S, A, L, R]

Matrix all rules:
[00, 21, 23]
[03, , ]
[(, , ]
[), , ]

Matrix T rules:
[]
[]
[(]
[)]

Matrix NT rules:
[00, 21, 23]
[03, , ]
```

Then the results, counter and runtime in *ms* is shown for each parsing method.

For the `BottomUp` method is the CYK algorithm table printed.

References

- [1] *Chomsky's Normal Form (CNF)*. Website. <https://www.javatpoint.com/automata-chomskys-normal-form>, opened on 26.09.2022.
- [2] *Dynamic Programming*. Website. url=<https://www.programiz.com/dsa/dynamic-programming>, opened on 26.09.2022.
- [3] Robert Eisele. *A CYK Algorithm Visualization*. May 2018. URL: <https://www.xarg.org/tools/cyk-algorithm/>.
- [4] Glenn K. Manacher. "An improved version of the Cocke-Younger-Kasami algorithm". In: *Computer Languages* 3.2 (1978), pp. 127–133. ISSN: 0096-0551. DOI: [https://doi.org/10.1016/0096-0551\(78\)90029-2](https://doi.org/10.1016/0096-0551(78)90029-2). URL: <https://www.sciencedirect.com/science/article/pii/0096055178900292>.
- [5] A.J. Kfoury Robert N. Moll Michael A. Arbib. *An Introduction to Formal Language Theory*. Springer-Verlag, 1988.
- [6] Fabio Massimo Zanzotto, Giorgio Satta, and Giordano Cristini. "CYK Parsing over Distributed Representations". In: *Algorithms* 13 (Oct. 2020), p. 262. DOI: 10.3390/a13100262.